Modelo predictivo de la Actividad Vegetal en el Bosque Tropical Seco del Cañón del Río Cauca

Eileen Melissa Arévalo Garnica
Pablo Uribe Uribe

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
David Manuel Villanueva Valdes, M.Sc

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2024

Cita	(Arévalo Garnica & Uribe Uribe, 2024)
Referencia **Estilo APA 7 (2020)**	Arévalo Garnica, E.M, & Uribe Uribe, P. (2024). *Modelo predictivo de la Actividad Vegetal en el Bosque Tropical Seco del Cañón del Río Cauca*. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VI.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Dedicatoria

Eileen Melissa Arévalo Garnica

A mi familia, por su cariño incondicional y por estar siempre a mi lado, apoyándome en cada paso que doy. A Lina, por su bonita compañía y por nutrirme mientras estudiaba. A Choco, por ser mi fiel compañero en las noches en vela.

Con todo mi amor y gratitud.

Agradecimientos

Queremos expresar nuestro agradecimiento a todas las personas que hicieron posible la realización de esta monografía.

En primer lugar, agradecemos a Arbei Osorio, estudiante de doctorado, por su orientación, por proporcionarnos la bibliografía esencial y por facilitarnos los datos con los que se ha realizado este estudio. Su colaboración ha sido fundamental para el éxito de nuestro trabajo.

A nuestro asesor el profesor David Manuel Villanueva por su guía y apoyo constante durante el proyecto.

Tabla de contenido

Lista de tablas

Lista de figuras

Siglas, acrónimos y abreviaturas

ARIMA	AutoRegressive Integrated Moving Average
CNN	Convolutional Neural Network (Red Neuronal Convolucional)
RNN	Recurrent Neural Network (Red Neuronal Recurrente)
LSTM	Long Short-Term Memory
NDVI	Normalized Difference Vegetation Index (Índice de Vegetación de Diferencia Normalizada)
IDEAM	Instituto de Hidrología, Meteorología y Estudios Ambientales
mm	Milímetros
MSE	Mean Squared Error (Error Cuadrático Medio)
MAE	Mean Absolute Error (Error Medio Absoluto)
RMSE	Root Mean Squared Error (Raíz del Error Cuadrático Medio)
R^2	Coeficiente de Determinación
CRISP-DM	Cross-Industry Standard Process for Data Mining (Proceso Estándar Intersectorial para la Minería de Datos)
PIRAGUA	Plan Integral de Recursos Ambientales y Gestión del Agua
POMCAS	Planes de Ordenación y Manejo de Cuencas Hidrográficas
MDE	Modelo Digital de Elevaciones
IGAC	Instituto Geográfico Agustín Codazzi
GPU	Graphics Processing Unit (Unidad de Procesamiento Gráfico)
TPU	Tensor Processing Unit (Unidad de Procesamiento de Tensor)
API	Application Programming Interface (Interfaz de Programación de Aplicaciones)

Resumen

Este proyecto se enfoca en evaluar las relaciones que puedan existir entre la actividad vegetal y las variables eco-hidrológicas para la creación de un modelo predictivo de la actividad vegetal en el bosque seco tropical del Cañón del Río Cauca en Colombia. La creación de una herramienta que ayude a predecir la actividad vegetal podría ser fundamental para hacer seguimiento de la salud del bosque en cuestión, para la gestión de los recursos hídricos y la conservación de estos ecosistemas. Para abordar este desafío, el proyecto utilizará métodos estadísticos, así como modelos de Deep Learning, y usará métricas de evaluación con el fin de comparar el comportamiento de estos y así encontrar el modelo más preciso y eficaz, utilizando como entrada variables eco-hidrológicas y ambientales, obtenidas de estaciones del IDEAM y de fotos satelitales de Google Earth Engine, logrando predecir la actividad vegetal (NDVI) en esta región. Aunque el alcance del proyecto se limita a esta cuenca específica y sus datos, su impacto potencial es relevante para las autoridades ambientales, los investigadores y las comunidades locales que buscan tomar decisiones informadas y promover la conservación de los bosques secos tropicales en la región.

Palabras clave: Bosque Tropical Seco, Actividad Vegeta, Predicción, NDVI (Índice de Vegetación de Diferencia Normalizada), Modelos de Deep Learning, Recursos Hídricos, Conservación de Ecosistemas, Series Temporales

https://github.com/emarevalog/Data_Science_Forecast_NDVI_Project

Abstract

This project focuses on evaluating the relationships that may exist between vegetation activity and eco-hydrological variables to create a predictive model of vegetation activity in the tropical dry forest of the Cauca River Canyon in Colombia. Developing a tool that helps predict vegetation activity could be crucial for monitoring the health of the forest, managing water resources, and conserving these ecosystems. To address this challenge, the project will use statistical methods as well as deep learning models and employ evaluation metrics to compare their performance and find the most accurate and effective model. The input will include eco-hydrological and environmental variables obtained from IDEAM stations and satellite images from Google Earth Engine, aiming to predict vegetation activity (NDVI) in this region. Although the project's scope is limited to this specific watershed and its data, its potential impact is significant for environmental authorities, researchers, and local communities seeking to make informed decisions and promote the conservation of tropical dry forests in the region.

Keywords: Tropical Dry Forest, Vegetation Activity, Prediction, NDVI (Normalized Difference Vegetation Index), Deep Learning Models, Water Resources, Ecosystem Conservation, Time Series

1. Descripción del problema

1.1. Problema de negocio

El objetivo de este estudio es pronosticar la actividad vegetal en el bosque seco tropical de la cuenca del río Cauca, en el tramo comprendido entre los municipios de La Pintada y Valdivia, Colombia. Este desafío implica realizar un análisis que nos lleve a comprender la relación entre la vegetación y los recursos hídricos, así como evaluar cómo los cambios en estos sistemas pueden influir en la conservación de los bosques secos tropicales y en la gestión de los recursos hídricos.

El problema surge debido a la intensificación de las actividades antropogénicas en la región, como la agricultura, la ganadería, la minería, el desarrollo urbano y el turismo. Estas actividades han llevado a la fragmentación de los bosques secos tropicales, poniendo en riesgo su futuro y los servicios ambientales que proporcionan.(Miles et al., 2006; Ministerio del Ambiente, 2017). El aumento en la frecuencia e intensidad de los incendios forestales, atribuido en gran medida a las disminuciones en las precipitaciones y a los períodos prolongados de sequía relacionados con el cambio climático, agrava aún más la situación. La falta de herramientas efectivas para medir el impacto o riesgo futuro de estas actividades en el ecosistema y en la cuenca del río Cauca resalta la necesidad de desarrollar una herramienta que permita a las autoridades y comunidades medir y mitigar esta problemática (Prance, 2006; Sánchez-Azofeifa & Portillo-Quintero, 2011).

Los bosques secos tropicales, se ubican en zonas bajas de regiones tropicales caracterizadas por climas cálidos durante todo el año y largas estaciones secas, representan una parte significativa de los ecosistemas naturales, especialmente en América Latina. Son cruciales debido a la amplia gama de servicios ambientales que proporcionan y la rica biodiversidad que albergan. Además, cumplen un papel esencial en la protección contra la erosión eólica e hídrica. A pesar de su importancia, los bosques secos tropicales enfrentan varios peligros significativos. La deforestación debido a la extracción de madera y la expansión agrícola representa una

amenaza constante para su preservación. El aumento de incendios forestales en las últimas décadas, debido a la disminución de precipitaciones y a los períodos prolongados de sequía relacionados con el cambio climático, es una de las mayores amenazas (Schröder et al., 2021).

Existe una relación estrecha entre la actividad vegetal y las variables climáticas. Los análisis de correlación han revelado que la temperatura ejerce un impacto negativo en la vegetación en regiones áridas y semiáridas, mientras que tiene un efecto positivo en el crecimiento de la vegetación en zonas húmedas a lo largo de toda la temporada de crecimiento. (Muradyan et al., 2019). Además, se ha observado una relación bidireccional entre la vegetación y las fuentes de agua. La disponibilidad de agua desempeña un papel fundamental en la salud de los ecosistemas y en los procesos esenciales para mantener las condiciones ecológicas y los servicios ecosistémicos.

La vegetación utiliza diversas fuentes de agua, incluyendo las superficiales y las aguas subterráneas. El nivel freático puede fluctuar debido a la absorción de agua por parte de las plantas, disminuyendo durante el día e incrementando durante la noche, cuando cesa la actividad de evapotranspiración de las plantas. Aunque el agua subterránea representa aproximadamente el 37% del suministro de agua para la vegetación, su importancia aún es poco comprendida y estudiada. (Wang et al., 2023).

Para estudiar las interacciones entre la actividad vegetal y las variables climáticas se pueden realizar modelado eco-hidrológico. Los modelos basados en productos de teledetección detectan cambios en las características aéreas de la vegetación, principalmente utilizando el Índice de Vegetación de Diferencia Normalizada (NDVI) (Wang et al., 2023). Los índices de vegetación desempeñan un papel importante en la evaluación del crecimiento y salud de la vegetación, lo que es de gran relevancia en diversas aplicaciones, especialmente en el seguimiento de cambios en la cobertura terrestre a partir de imágenes satelitales.

Los índices de vegetación son expresiones matemáticas que integran mediciones de reflectancia en las bandas espectrales. Uno de los índices más ampliamente reconocidos y

utilizados es el NDVI, que proporciona valores que oscilan entre -1 y 1. En este contexto, los valores positivos indican un aumento en la densidad y vigor de la vegetación, mientras que los valores negativos señalan áreas carentes de cobertura vegetal (Ferchichi et al., 2022).

En Colombia, el conocimiento sobre el bosque seco tropical es limitado y carece de datos sólidos que permitan una gestión integral de este importante ecosistema. La distribución de este tipo de bosque en Colombia está estrechamente relacionada con procesos históricos de deforestación y colonización. En la actualidad, varios factores ejercen una gran presión sobre estos ecosistemas estratégicos, incluyendo la agricultura, la ganadería, la minería, el desarrollo urbano y el turismo. Esta presión ha llevado a la fragmentación de los bosques secos tropicales, lo que representa una amenaza para su conservación y para los servicios ecosistémicos que proporcionan. Es crucial abordar esta problemática a través de una gestión adecuada y la generación de información sólida que respalde la toma de decisiones en la conservación de estos ecosistemas (Pizano & Garcia, 2014).

El área ribereña del río Cauca en el departamento de Antioquia, Colombia, ha sido objeto de una explotación intensiva de sus recursos naturales a lo largo de la historia, desde la época de la conquista y colonización española hasta la actualidad. Esta explotación ha involucrado actividades como la minería, la tala de árboles y la pesca. Como resultado de estas actividades, la proporción de bosque seco tropical en la región, tanto en el Alto Cauca como en el Medio y Bajo Cauca, ha disminuido significativamente y actualmente representa menos del 20% de la cobertura total. Este proceso de fragmentación ha dado lugar a la presencia de parches de bosque seco tropical dispersos en un paisaje transformado por actividades humanas (Pizano & Garcia, 2014).

1.2. Aproximación desde la analítica de datos

Los modelos predictivos desarrollados en este proyecto buscan anticipar y comprender la actividad vegetal en los bosques secos tropicales del Cañón del Río Cauca. Estos modelos se centran en la relación entre las variables eco-hidrológicas, como la precipitación, humedad del suelo, evapotranspiración, recarga, y la actividad vegetal, medida mediante el Índice de Vegetación de Diferencia Normalizada (NDVI).(Nemani et al., 2003; Turner et al., 2003).

Su función principal es ofrecer un sistema de monitoreo continuo de la actividad vegetal en la región, permitiendo la detección temprana de cambios en la salud de los bosques. Emiten alertas tempranas sobre posibles amenazas, como incendios forestales o sequías prolongadas, para facilitar respuestas proactivas. Además, respaldan la toma de decisiones informada para autoridades ambientales, investigadores y comunidades locales al proporcionar una forma numérica y gráfica de cómo las variables eco-hidrológicas impactan la vegetación.(Vicente-Serrano et al., 2013)

Estos modelos contribuyen a la planificación sostenible al ofrecer información valiosa sobre la relación entre las actividades humanas, el cambio climático y la salud de los bosques. Ayudan a desarrollar políticas y prácticas que promuevan la conservación a largo plazo de estos ecosistemas.(Scanlon et al., 2007). Además, podrían desempeñar un papel importante en la gestión de recursos hídricos, al proporcionar entendimiento sobre cómo las variables ambientales afectan la vegetación y mantener el equilibrio hídrico en la cuenca del Río Cauca para un uso sostenible del agua.

En resumen, estos modelos no solo ofrecen una visión profunda de las interacciones eco-hidrológicas en los bosques secos tropicales, sino que también brindan herramientas prácticas para abordar desafíos específicos, proteger la biodiversidad y promover la gestión sostenible de estos valiosos ecosistemas.

1.3. Origen de los datos

El conjunto de datos utilizado en este estudio proviene de diversas fuentes, proporcionando una perspectiva integral para el desarrollo del modelo predictivo. En primer lugar, se emplean imágenes satelitales de alta resolución accesibles a través de la plataforma Google Earth Engine[1], obtenidas de satélites como Landsat y Sentinel[2]. Estas imágenes, capturadas a lo largo del tiempo, ofrecen una visión detallada de la cuenca de estudio y son fundamentales para rastrear cambios en la actividad vegetal.

[1] Google Earth Engine. Disponible en: https://earthengine.google.com/
[2] Landsat. Disponible en: https://landsat.usgs.gov/; Sentinel. Disponible en: https://sentinel.esa.int/

El Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM)[3] contribuye significativamente al proyecto mediante datos de sus estaciones de monitoreo distribuidas en la región. Estos datos, que abarcan parámetros hidroclimáticos como precipitación, evapotranspiración y humedad del suelo, proporcionan una comprensión detallada de las condiciones ambientales locales.

La información topográfica y geoespacial se integra para contextualizar el entorno del estudio. Datos topográficos y mapas de cobertura terrestre enriquecen la comprensión de las características del paisaje, complementando así el análisis de las variables eco-hidrológicas.

Además, se incluyen datos recopilados durante una investigación doctoral, que abarcan mediciones de campo y análisis de suelos. Estos datos aportan una perspectiva única y específica relacionada con la investigación en curso y son propiedad del autor de la investigación doctoral, por lo que no están disponibles públicamente.

Finalmente, la revisión de la bibliografía científica y estudios previos sobre la ecología de los bosques secos tropicales agrega profundidad al conjunto de datos. Esta revisión complementaria asegura que el proyecto se beneficie de los conocimientos acumulados en el campo, fortaleciendo así las bases teóricas y prácticas del modelo predictivo.

Estas diversas fuentes de datos, al combinarse, proporcionan una base sólida y multifacética para el desarrollo de modelos predictivos precisos y robustos, permitiendo una evaluación detallada y fiable de la actividad vegetal en la cuenca del río Cauca.

1.4. Métricas de desempeño

Durante el desarrollo de nuestro modelo de predicción, en la etapa de experimentación donde el objetivo es seleccionar el mejor tipo de modelo y, además, los mejores hiperparámetros para alcanzar el mayor nivel de precisión en las predicciones, utilizaremos cuatro métricas fundamentales para comparar cada combinación de modelo y sus hiperparámetros. La eficacia de las técnicas de series de tiempo se puede evaluar utilizando métricas como el error medio

[3] Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). Disponible en: http://www.ideam.gov.co/

absoluto (MAE), el error porcentual absoluto medio (MAPE), el error cuadrático medio (MSE) y el error cuadrático medio (RMSE), ya que estas métricas permiten medir la precisión y exactitud de las predicciones realizadas por los modelos. Estas métricas son ampliamente utilizadas en la literatura para evaluar el desempeño de los modelos de predicción (Wang, Zhang, & Li, 2024).

- **MAE (Media Absoluta de Errores):** Mide la media de las diferencias absolutas entre los valores observados y los valores predichos. Un valor más bajo indica un mejor ajuste del modelo a los datos.
- **MSE (Media de los Errores al Cuadrado):** Mide las diferencias entre los valores observados y los valores predichos al cuadrado. Al elevar al cuadrado los errores, magnifica los errores grandes, por lo que hay que utilizarla con cuidado cuando se tienen valores anómalos en el conjunto de datos.
- **RMSE (Error Cuadrático Medio):** Mide la raíz cuadrada de la diferencia entre los valores observados y los valores predichos. Un valor más bajo de RMSE indica un mejor ajuste del modelo a los datos.
- **R^2 (Coeficiente de Determinación):** Evalúa la proporción de la varianza en la variable de respuesta que es explicada por el modelo. Un valor más alto de R^2 indica una mejor capacidad del modelo para explicar la variabilidad en los datos.

En nuestro alcance no está llevar el modelo a producción ni a ser usado por las entidades que gestionan la regulación para el cuidado de estos ecosistema, pero como métricas de negocio a usar en un futuro recomendamos, tener mensualmente un informe con:

- **Coeficiente de Determinación (R^2) mes vencido**: este indicador nos podrá mostrar la precisión del modelo en ese momento del tiempo y tomar acciones como reentrenar, cambiar o cancelar el modelo.
- **R^2/Costo mensual**: Este modelo deberá estar corriendo y consumiendo algunos recursos que pueden ser costeados, una relación entre el R^2 y este costo podría ser un valor importante para decidir si vale la pena seguir con el modelo corriendo.
- **Efectividad de las acciones tomadas**: Las entidades encargadas deberían de poder evidenciar la efectividad de sus acciones, y una vez ejecutada una intervención o aprobada e impuesta una ley, el modelo debería mostrar cómo sus predicciones se tornan hacia una actividad vegetal positiva.

- **Comparación de Resultados del Modelo con Variaciones del NDVI**: Este indicador nos permite comprobar la consistencia entre las predicciones del modelo y los datos observados de NDVI.

2. Objetivos

2.1. Objetivo general

Desarrollar y evaluar modelos predictivos para pronosticar la actividad vegetal en el bosque seco tropical de la cuenca del río Cauca, en el tramo comprendido entre los municipios de La Pintada y Valdivia, Colombia. Estos modelos buscan comprender la relación entre la vegetación y los recursos hídricos, así como evaluar cómo los cambios en estos sistemas pueden influir en la conservación de los bosques secos tropicales y en la gestión de los recursos hídricos.

2.2. Objetivos específicos

- Recopilar y preprocesar datos de imágenes satelitales, estaciones de monitoreo del IDEAM, información topográfica, datos de investigación doctoral y bibliografía científica. Realizar un proceso de limpieza, eliminación de valores atípicos y normalización para garantizar la calidad y coherencia de los datos.
- Identificar relaciones entre las variables eco-hidrológicas y el NDVI mediante un análisis exploratorio de datos para analizar posibles correlaciones y patrones que sirvan como base para el desarrollo del modelo predictivo.
- Seleccionar variables relevantes utilizando técnicas de selección para identificar y evaluar la importancia de cada variable y su contribución al modelo en la predicción de la actividad vegetal.
- Implementar y evaluar modelos ARIMA y SARIMAX para capturar las tendencias y patrones estacionales en la serie temporal de NDVI.

- Desarrollar y optimizar modelos de deep learning, como Redes Neuronales Convolucionales (CNN), Redes Neuronales Recurrentes (RNN) y Redes de Memoria a Largo Plazo (LSTM), para mejorar la precisión en el pronóstico del NDVI..
- Comparar el desempeño de los modelos ARIMA/SARIMAX y los modelos de deep learning utilizando métricas de rendimiento como MAE, MSE, RMSE y R^2 para determinar el modelo más efectivo en la predicción de la actividad vegetal..
- Seleccionar y optimizar el modelo con mejores métricas de rendimiento, para su futura entrega y publicación.

3. Datos

3.1. Datos originales

La naturaleza estructural de nuestra fuente de datos es el resultado de una investigación doctoral previa, en la cual se llevó a cabo un riguroso trabajo de preparación y construcción del dataset. En la siguiente tabla, se presenta una descripción detallada de los datos originales utilizados en dicha investigación, los cuales serán la base para nuestro modelo.

Tabla 1

Datos originales investigación doctoral "Dinámica ecohidrológica entre bosque seco tropical y agua subterránea en el cañón del río Cauca en la región comprendida entre los municipios de Caramanta y Valdivia, Antioquia-Colombia"

Información	Descripción
Datos hidrometeorológicos de la red de monitoreo del IDEAM	Datos de precipitación total, temperatura máxima, temperatura mínima, temperatura media, evaporación, caudal medio, humedad relativa y brillo solar
Datos hidrometeorológicos de la red de monitoreo de PIRAGUA-CORANTIOQUIA	Datos de precipitación total, temperatura máxima, temperatura mínima, temperatura media, evaporación, caudal medio, humedad relativa y brillo solar
Unidades cartográficas de suelos – escala 1:100.000	Shape de las unidades cartográficas de suelos del Estudio General de Suelos y Zonificación de Tierras del departamento de Antioquia (IGAC & Gobernación de Antioquia, 2007)
Información textural y de propiedades hidráulicas de suelos – escala 1:100.000	Información de los perfiles de suelos del Estudio General de Suelos y Zonificación de Tierras de Antioquia (IGAC & Gobernación de Antioquia, 2007)
Información de suelos de los	Descripción de suelos a escala 1:25.000 de los

POMCAS en jurisdicción de CORANTIOQUIA – escala 1:25.0000	POMCAS del río Amagá – quebrada Sinifaná y Río Aurrá
Información textural y de propiedades hidráulicas de suelos – escala 1:25.000	Información de los perfiles de suelos de los POMCAS del río Amagá – quebrada Sinifaná y Río Aurrá
Mapa de coberturas de la tierra a escala 1:100.000	Mapa de coberturas de la tierra para la zona de estudio con su respectiva leyenda – escala 1:100.000. Formato shape.
Mapa de coberturas de la tierra a escala 1:25.000 de los POMCAS en jurisdicción de CORANTIOQUIA	Mapa de coberturas de la tierra de los POMCAS del río Amagá – quebrada Sinifaná y Río Aurrá, a escala 1:25.000. Formato shape.
Modelo Digital de Elevaciones (MDE) para Antioquia	Modelo Digital de Elevaciones (MDE) para Antioquia, tamaño de píxel 12,5 m, obtenido de la base de datos ALOS – PALSAR y corregido por la Gobernación de Antioquia. Formato raster.
Modelo Digital de Elevaciones (MDE) de Colombia	Modelo Digital de Elevaciones (MDE) para Antioquia, tamaño de pixel 30m, obtenido de información satelital ASTER y corregido por IGAC. Formato raster.
Red de drenaje superficial	Red de drenaje superficial (drenaje doble y drenaje sencillo) en escala 1:100.000 y 1:25.000. Formato shape.

Nota:Tabla tomada de"Presentación 5-12-2022 SEMINARIO Geolimna" (Osorio Restrepo, 2023)

Como resultado del tratamiento de datos y del procesamiento de las imágenes satelitales realizado durante la investigación doctoral previa, se obtuvo un dataset preparado. Este dataset será utilizado como base en nuestro estudio, sobre el cual llevaremos a cabo un preprocesamiento adicional para adaptarlo a las necesidades específicas de nuestro modelo.

3.2. Dataset

El dataset de entrada al modelo es un conjunto de datos que contiene información climática mensual promedio desde el año 2013 hasta 2021 . La variable NDVI fue obtenida por procesamiento de imágenes del satélite MODIS y las demás provienen del modelo Soil-Water Balance (SWB). Las variables del dataset, todas medidas en milímetros excepto el NDVI que es adimensional, son las siguientes:

- Mes (1 al 12)
- NDVI
- Precipitación (mm)
- Evapotranspiración real (mm)
- Evapotranspiración potencial (mm)
- Intercepción (mm)
- Humedad del Suelo (mm)
- Recarga (mm)

De estas variables, hemos decidido utilizar tres: Precipitación, Humedad del Suelo y Recarga. Esta selección se justifica porque estas variables provienen de mediciones directas, mientras que las demás se obtienen a partir de cálculos realizados con modelos. Además, nuestro análisis exploratorio de datos, detallado en el capítulo 3.3, respalda esta elección al mostrar la fuerte relación de estas variables con el NDVI.

En el capítulo 4, se detallan los pasos adicionales que llevamos a cabo antes de utilizar estos datos en nuestro modelo. Primero, especificamos la frecuencia temporal de los datos (mensual) y verificamos la estacionariedad de la serie temporal utilizando la prueba Dickey-Fuller Aumentada (ADF). Al encontrar que los datos originales no son estacionarios, aplicamos diferenciación regular y estacional para lograr la estacionariedad necesaria.

Además, las variables exógenas (precipitación, evapotranspiración, intercepción, humedad del suelo y recarga) se escalan utilizando el método *MinMaxScaler*, que ajusta los valores dentro del rango [0,1]. Este método es adecuado para nuestros datos, que no siguen una distribución normal y están sesgados.

Finalmente, para los modelos de deep learning, escalamos los datos, creamos secuencias de entrada y salida, convertimos estas secuencias en tensores y dividimos el conjunto de datos en conjuntos de entrenamiento y prueba. Estos pasos aseguran que los datos estén preparados adecuadamente para ser procesados por los modelos.

3.3. Analítica descriptiva

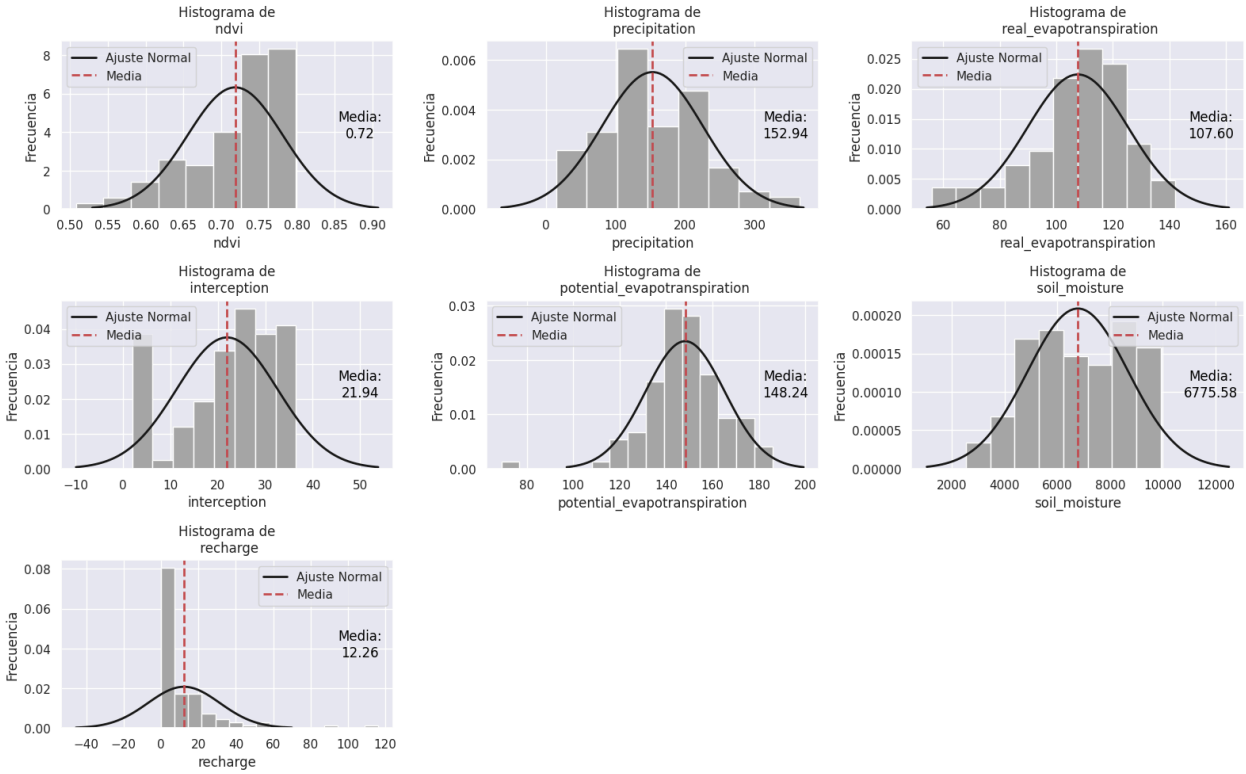
Con el objetivo de conocer a fondo el comportamiento y composición de nuestras variables hicimos un ejercicio de graficar diferentes escenarios para tener evidencia gráfica y realizar el análisis (Arévalo Garnica & Uribe Uribe, 2024).

La Figura 1 muestra los histogramas de las variables, en orden de izquierda a derecha: NDVI, Precipitación, Evapotranspiración, Intercepción, Evapotranspiración Potencial, Humedad del Suelo y Recarga, junto con sus ajustes normales correspondientes. La visualización de la distribución de estas variables es crucial en el análisis descriptivo previo a la construcción de modelos estadísticos o de Deep Learning. Esto permite comprender la naturaleza de los datos y detectar posibles patrones o anomalías.

- NDVI: Distribución con varios picos, sugiriendo valores modales múltiples o variabilidad estacional.
- Precipitación: Distribución ligeramente sesgada a la izquierda, con una media de aproximadamente 100 mm.
- Evapotranspiración: Distribución uniforme con un ligero sesgo a la derecha y una media de alrededor de 110 mm.
- Intercepción: Distribución con varios picos menores y una media cerca de 21.5 mm.
- Evapotranspiración Potencial: Distribución normal con una media de alrededor de 200 mm.
- Humedad del suelo: Distribución sesgada a la derecha con una media de 8175 mm.
- Recarga: Distribución bimodal con un ajuste normal menos preciso y una media de 16.4 mm.

Figura 1

Histogramas de variables relacionadas con el pronóstico del NDVI.

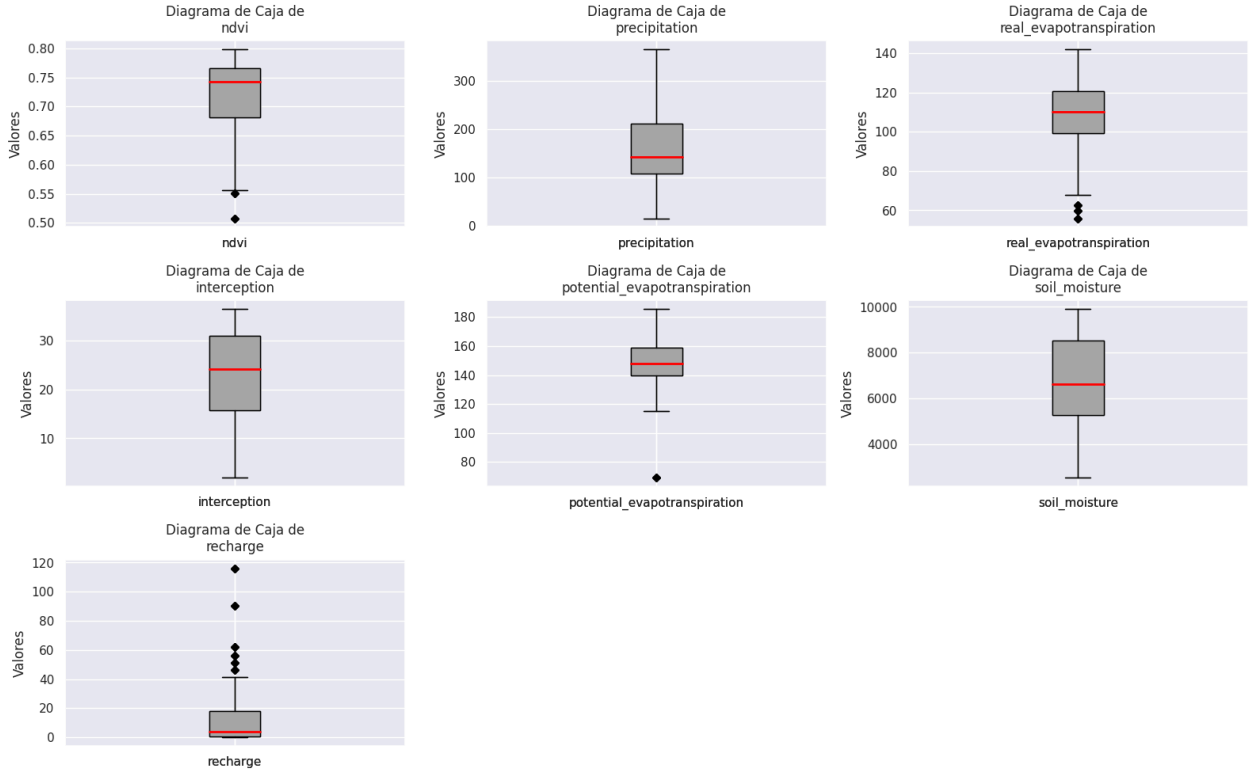


Con el fin de comprender mejor la distribución de los valores de nuestros datos y visualizar gráficamente cuántos y de qué magnitud son nuestros valores atípicos, decidimos realizar diagramas de caja y bigotes. Esto nos permite tomar decisiones informadas sobre si debemos conservar, transformar o eliminar estos valores atípicos.

La Figura 2 muestra que la distribución del NDVI tiene una variabilidad relativamente baja, con algunos valores atípicos que podrían representar cambios estacionales significativos. Las variables Precipitación, Intercepción, Evapotranspiración Potencial y Evapotranspiración Real presentan distribuciones uniformes sin valores atípicos significativos, lo que sugiere que estos procesos son bastante consistentes a lo largo del tiempo. La Humedad del Suelo muestra una amplia gama de valores, reflejando la variabilidad en la cantidad de agua retenida en el suelo, sin valores atípicos significativos. La variable Recarga muestra una variabilidad considerable y varios valores atípicos, lo que sugiere eventos de recarga inusuales o posibles inconsistencias en la medición.

Figura 2

Diagramas de caja para las variables NDVI, Precipitación, Evapotranspiración Real, Intercepción, Evapotranspiración Potencial, Humedad del Suelo y Recarga.



La evaluación de la dispersión y correlación de las variables frente al NDVI es crucial para comprender la variabilidad de estos parámetros y su interrelación. Esta información es esencial para evaluar la consistencia y estabilidad de los datos, especialmente en el contexto de la gestión de recursos hídricos.

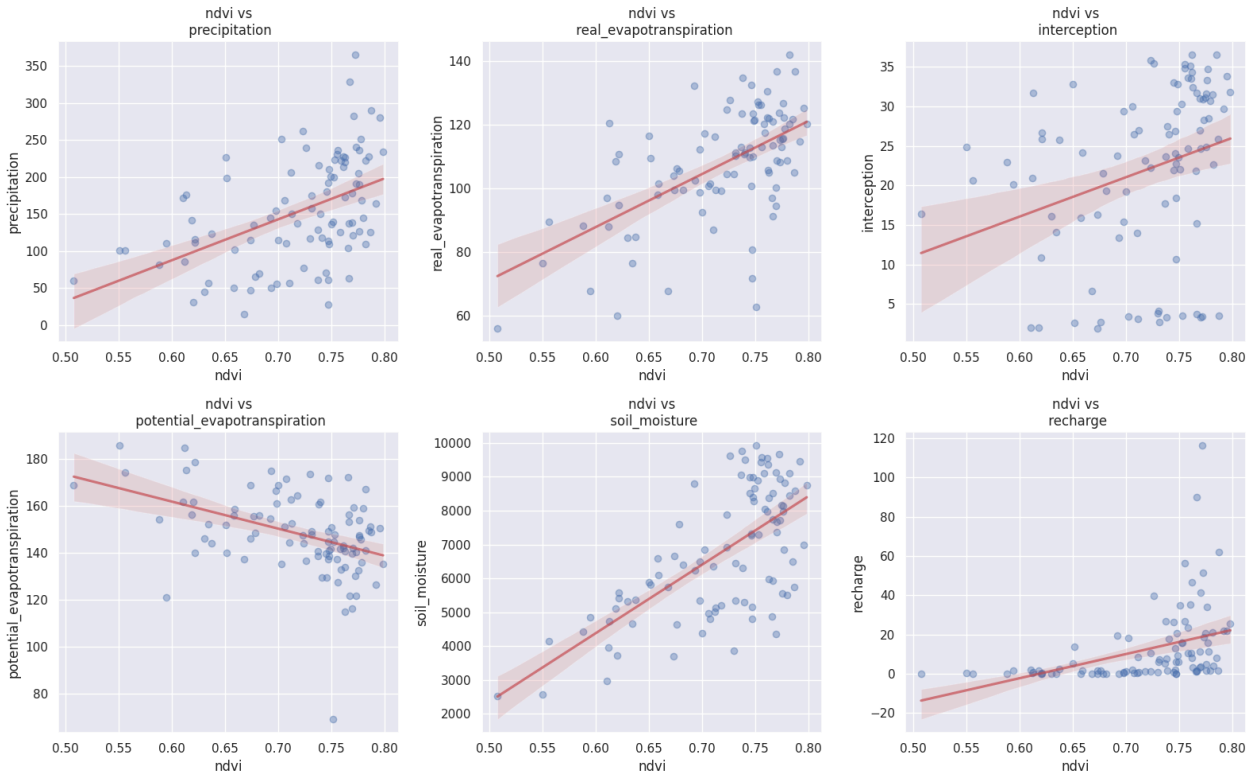
Los gráficos de la Figura 3 muestran la relación entre el NDVI y las demás variables. Cada gráfico incluye una línea de regresión con un intervalo de confianza sombreado, lo que ayuda a visualizar las tendencias y la fuerza de las relaciones entre las variables.

Observamos que la precipitación, la evapotranspiración real, la intercepción, la humedad del suelo y la recarga tienen relaciones positivas con el NDVI, indicando que mayores valores en estas variables generalmente se asocian con mayores niveles de vegetación. Por otro lado, la evapotranspiración potencial muestra una relación negativa con el NDVI, lo que sugiere un mayor estrés hídrico en condiciones de alta demanda evaporativa. La humedad del suelo presenta

una de las relaciones más fuertes con el NDVI, subrayando la importancia del agua disponible en el suelo para la vegetación.

Figura 3

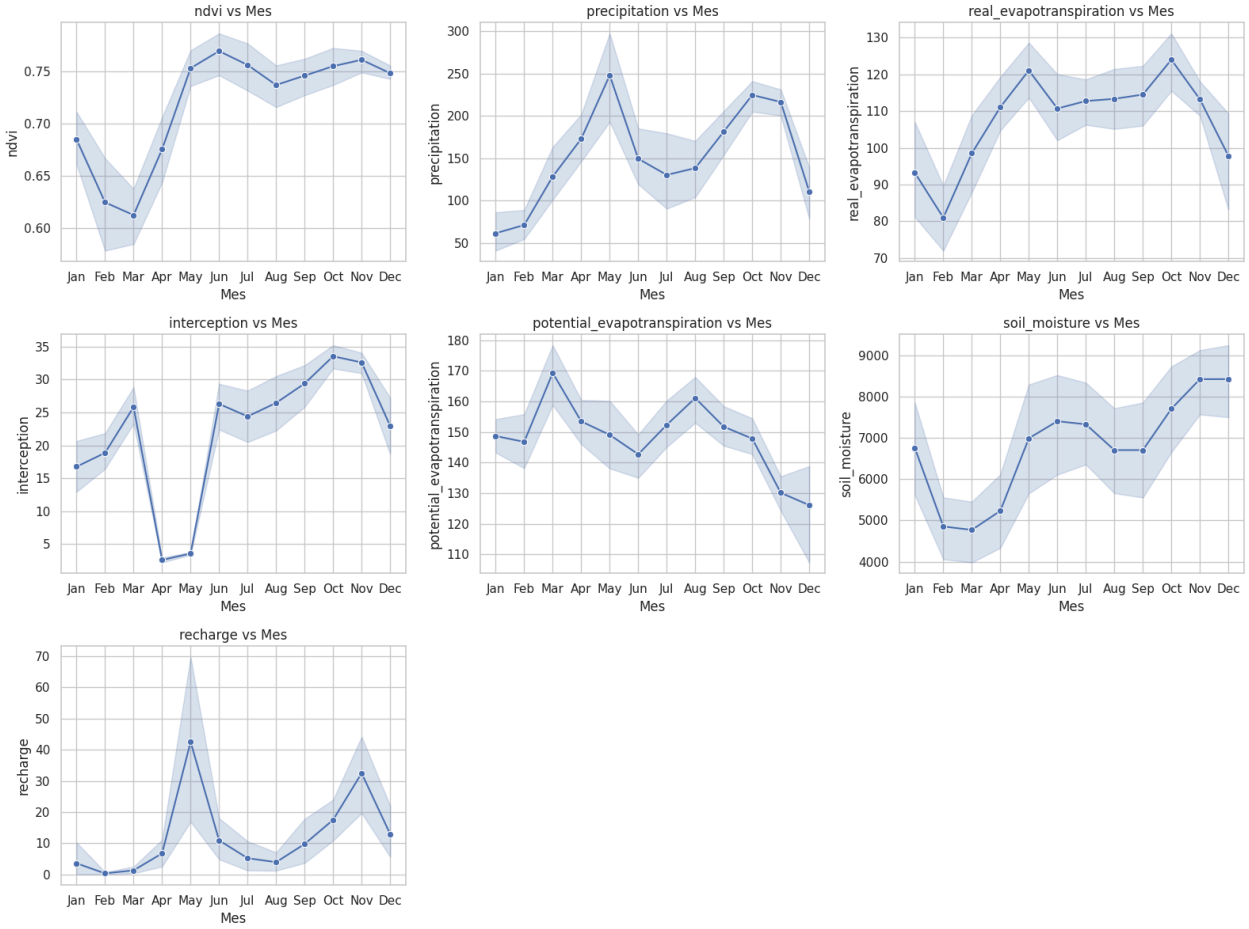
Gráficos de dispersión entre el NDVI y las variables Precipitación, Evapotranspiración Real, Intercepción, Evapotranspiración Potencial, Humedad del Suelo y Recarga.



Al analizar el ciclo anual de cada parámetro, podemos observar la variación y la tendencia de las mediciones a lo largo de los 8 años, desde 2013 hasta 2021. En la Figura 4 se presentan los gráficos de línea que muestran la evolución mensual de diversas variables relacionadas con el NDVI a lo largo del año. Los patrones estacionales son evidentes en todas las variables, con la mayoría de ellas mostrando picos en mayo. Esto resalta la importancia de la estacionalidad en la dinámica de la vegetación y la disponibilidad de agua. La relación entre la precipitación, la evapotranspiración y la humedad del suelo con el NDVI subraya la interdependencia de estas variables en el crecimiento y la salud de la vegetación.

Figura 4

Gráficos de línea de NDVI, Precipitación, Evapotranspiración Real, Intercepción, Evapotranspiración Potencial, Humedad del Suelo y Recarga vs Mes.



Por último, la matriz de correlación presentada en la Figura 5 muestra las correlaciones entre las variables. La correlación es una medida estadística que indica la relación entre dos variables, con valores que van de -1 (correlación negativa perfecta) a 1 (correlación positiva perfecta).

La matriz revela relaciones importantes:

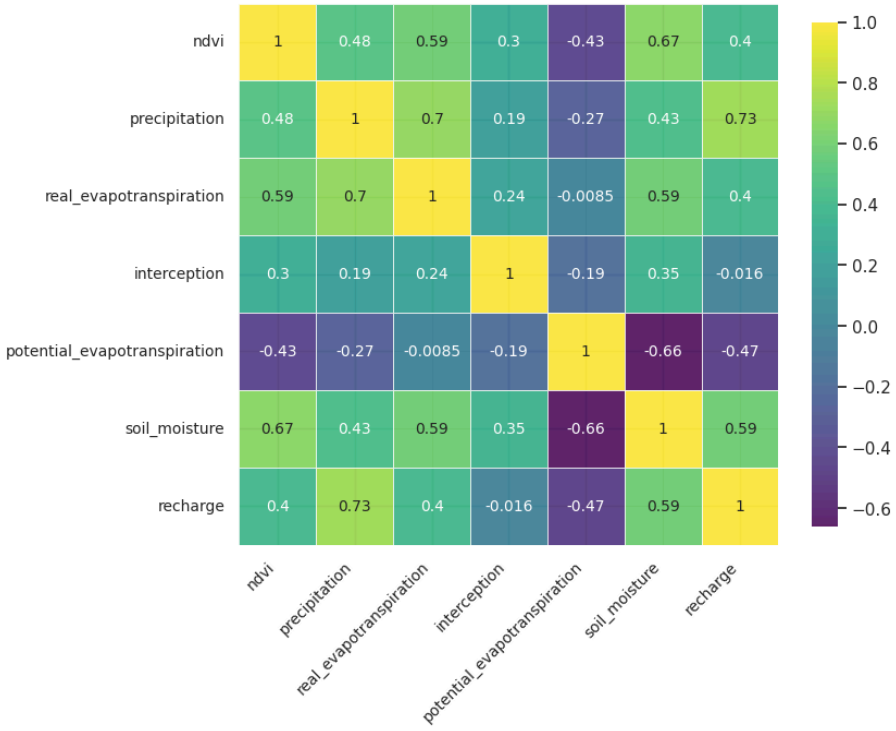
- Relaciones fuertes: Entre el NDVI y la humedad del suelo (0.67).

- Relaciones moderadas: Entre el NDVI y la precipitación (0.48), la evapotranspiración real (0.59) y la recarga (0.40), así como una correlación negativa moderada con la evapotranspiración potencial (-0.43).
- Relaciones débiles: Entre el NDVI y la intercepción (0.30).

Estas correlaciones indican cómo las diferentes variables ambientales influyen en el NDVI. Notamos la importancia de la disponibilidad de agua en el suelo y la precipitación para la vegetación..

Figura 5

Matriz de correlación entre las variables NDVI, Precipitación, Evapotranspiración Real, Intercepción, Evapotranspiración Potencial, Humedad del Suelo y Recarga.



4. Proceso de analítica

4.1. Pipeline principal

Figura 6

Diagrama del proceso CRISP-DM.

Nota: Los iconos utilizados en este diagrama fueron proporcionados por Freepik.

4.2. Preprocesamiento

4.2.1. Preprocesamiento de Datos para el Modelo ARIMA

- Especificación de la Frecuencia Temporal y Verificación de Estacionariedad: Se especifica la frecuencia temporal de los datos, que en nuestro caso es mensual. A

continuación, se verifica la estacionariedad de la serie temporal original utilizando la prueba Dickey-Fuller Aumentada (ADF), que tiene las siguientes hipótesis:

- Hipótesis Nula (H0): La serie temporal tiene una raíz unitaria (no es estacionaria).
- Hipótesis Alternativa (H1): La serie temporal no tiene una raíz unitaria (es estacionaria).

- Los resultados de la prueba ADF antes de las transformaciones sugieren que los datos no son estacionarios. Para que una serie temporal sea estacionaria, el valor p debe ser menor que el nivel de significancia (comúnmente 0.05). En nuestro caso, tenemos:
  - ADF Statistic: -1.827
  - p-value: 0.367

- Dado que el valor p (0.367) es mayor que 0.05, no podemos rechazar la hipótesis nula de que la serie tiene una raíz unitaria, lo que indica que los datos no son estacionarios. Por lo tanto, debemos realizar transformaciones para que los datos sean estacionarios antes de aplicar el modelo ARIMA.

- Transformaciones para Lograr Estacionariedad: Aplicamos la Diferenciación Regular y la Diferenciación Estacional para eliminar tendencias lineales y patrones estacionales, respectivamente. Después de cada una de las diferenciaciones, se vuelve a aplicar la prueba ADF, cuyos resultados se presentan en la Tabla 2 (Arévalo Garnica & Uribe Uribe, 2024b).

Tabla 2

Métricas de evaluación para determinar estacionalidad de datos

	Después de Diferenciación Regular	Después de Diferenciación Estacional

ADF Statistic: -9.085	ADF Statistic: -3.968
p-value: 3.959e-15 (prácticamente 0)	p-value: 0.0016

Nota: Los resultados fueron obtenidos del análisis realizado por Arévalo Garnica & Uribe Uribe (2024b), cuyo código está disponible en [GitHub](https://github.com/emarevalog/Data_Science_Forecast_NDVI_Project/blob/main/ARIMA_model.ipynb).

Estos resultados muestran que, después de la diferenciación regular, la serie es estacionaria, ya que el valor p es mucho menor que 0.05. Esto indica que no hay una tendencia significativa en los datos diferenciados. Además, después de la diferenciación estacional, la serie también es estacionaria, ya que el valor p es menor que 0.05. Por lo tanto, las transformaciones aplicadas han sido efectivas para hacer que la serie temporal sea estacionaria, lo que nos permite proceder con la aplicación del modelo ARIMA.

- Escalado de Variables Exógenas: Las variables (precipitation, soil_moisture, recharge) se escalan entre [0,1] utilizando MinMaxScaler, dado que nuestros datos no tienen distribución normal y están sesgados, este método no asume una distribución específica de los datos y puede manejar mejor las distribuciones sesgadas al escalar los valores dentro de un rango específico .
  - $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$
    - x *es el valor original de los datos.*
    - x_{min} *es el valor mínimo del conjunto de datos.*
    - x_{min} *es el valor mínimo del conjunto de datos.*

4.2.2. Preprocesamiento de Datos para los Modelos de Deep Learning

- Escalado de datos con *MinMaxScale*r: Se aplicó el escalado de los datos utilizando el método *MinMaxScaler* para normalizar las características y asegurarse de que todos los

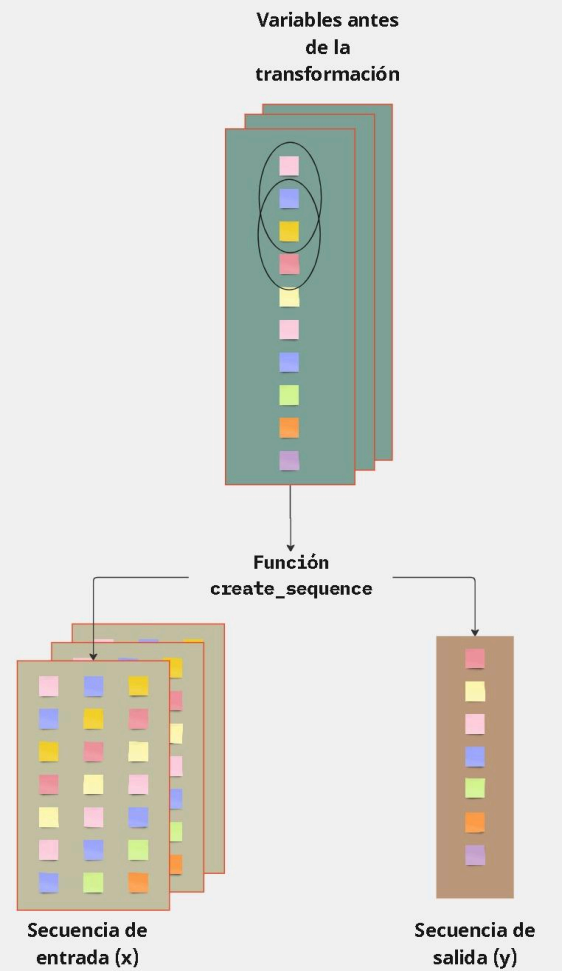
valores se encuentren dentro de un rango específico, mejorando así la eficiencia y precisión del entrenamiento de los modelos.

- Creación de secuencias con la función *create_sequences*: Esta función transforma los datos en secuencias adecuadas para el entrenamiento de modelos de *Deep Learning*. Toma un conjunto de datos y los convierte en secuencias de entrada y salida. La figura 7, ilustra el antes y el después de las transformaciones que tienen cada una de las variables.
    - Secuencia de entrada (X): Es un array de *NumPy* con tres dimensiones (3 variables usadas) de la forma (*n_secuencias, n_past_steps, n_features*), donde *n_secuencias* es el número total de secuencias generadas, *n_past_steps* es el número de pasos anteriores especificados, y *n_features* es el número de características en los datos (si *include_target_as_feature* es falso, *n_features* será el número de columnas menos una).
    - Secuencia de salida (y): Es un *array* de *NumPy* con dos dimensiones de la forma (*n_secuencias, n_forecast_steps*), donde *n_forecast_steps* es el número de pasos futuros a predecir.

      Esta función permite preparar de manera eficiente los datos para el entrenamiento de modelos de deep learning, asegurando que las secuencias de entrada y salida estén correctamente alineadas para la predicción de series temporales, como las Redes Neuronales Recurrentes (RNN) y las Redes Neuronales de Memoria a Largo Plazo (LSTM).

Figura 7

Ilustración de la creación de secuencias con la función create_sequences para una sola característica o variable



- Conversión a Tensores: Se convierten las secuencias de entrada y salida a tensores de *TensorFlow* para ser utilizados en el entrenamiento de modelos.

- División de Datos: Se divide el conjunto de datos en conjuntos de entrenamiento (80%) y prueba (20%).

- Al organizar los datos en una estructura tridimensional, creamos un enfoque de ventana deslizante que permite al modelo capturar dependencias y patrones temporales dentro de un contexto mensual. Con un paso de tiempo de 1, nuestro objetivo es capturar variaciones y relaciones mensuales en los datos del índice de vegetación NDVI (Smith et al., 2024).

- Evaluación de Modelos para Encontrar el Mejor Look Back, Número de Épocas y Batch Size: El proceso de evaluación de modelos se diseñó para identificar la mejor combinación de *look_back*, epochs y batch_size para predecir el índice de vegetación NDVI utilizando diferentes arquitecturas de redes neuronales (CNN, RNN y LSTM). A continuación, se describe el procedimiento seguido para esta evaluación:
    - Definición de Parámetros:
        - look_back_values: Valores de pasos hacia atrás utilizados en la secuencia (1, 2, 3, ..., 10).
        - epochs_values: Cantidad de épocas para el entrenamiento del modelo (100, 200, 300).
        - batch_size_values: Tamaño del batch para el entrenamiento (16, 32).
        - model_types: Tipos de modelos a evaluar (CNN, RNN, LSTM).

Los hiperparámetros fueron elegidos teniendo en cuenta el mejor RMSE se observan en la siguiente tabla:

Tabla 3

Mejores hiperparametros por modelo

Model	Look_Back	Epochs	Batch_Size	MAE	MSE	RMSE	R^2
CNN	6	100	32	0.003582	0.049736	0.003582	0.868009
RNN	6	300	32	0.006531	0.006531	0.006531	0.759330
LSTM	6	300	32	0.004145	0.047683	0.004145	0.847251

Nota: Los resultados fueron obtenidos del análisis realizado por Arévalo Garnica & Uribe Uribe (2024d), cuyo código está disponible en [GitHub](https://github.com/emarevalog/Data_Science_Forecast_NDVI_Project/blob/main/Test_Hiperatameters_DL.ipynb).

Sin embargo, luego de entrenar los modelos y ajustar hiperparámetros todos los modelos mostraron mejores resultados con Batch_Size igual a 5.

4.3. Modelos

4.3.1. ARIMA (Autoregressive Integrated Moving Average)

Un modelo ARIMA predice un valor en una serie temporal como una combinación lineal de sus propios valores pasados, es decir, el valor futuro que buscamos predecir depende de los valores anteriores de la misma serie. Este modelo puede utilizar la información del promedio móvil y los errores de predicción anteriores para mejorar las predicciones futuras (Wang, Zhang, & Li, 2024). Un modelo ARIMA se denomina modelo ARIMA(p,d,q), donde p es el número de términos autorregresivos (observaciones rezagadas), d es el número de diferencias necesarias para hacer la serie estacionaria, y q es el número de términos de media móvil (errores de pronóstico rezagados) en la predicción. La estacionariedad se verifica mediante la prueba de Dickey–Fuller, que prueba la hipótesis nula de que la serie temporal no es estacionaria. Si el valor p devuelto por la prueba es mayor que 0.05, la serie temporal no es estacionaria. Para hacer una serie estacionaria, es necesario "diferenciar" los datos (Bouznad et al., 2020).

Un modelo ARIMA también puede ser parametrizado y calibrado automáticamente utilizando la función Auto ARIMA, que facilita la parametrización automática del modelo, optimizando su precisión mediante criterios estadísticos como AIC y RMSE (Kesavan et al., 2021). Sin embargo, cuando la serie temporal exhibe estacionalidad, las tendencias estacionales pueden considerarse agregando parámetros adicionales.

El modelo ARIMA se ha aplicado con éxito en diversos estudios para pronosticar variables ambientales clave. Por ejemplo, Kesavan et al. (2021) utilizaron el modelo ARIMA para pronosticar la temperatura de la superficie terrestre y determinar la isla de calor urbana en la ciudad de Chennai, India. Esta metodología es relevante para el análisis del NDVI, ya que ambos parámetros están relacionados con la salud y el comportamiento de la vegetación. Al pronosticar el NDVI utilizando modelos ARIMA, se puede obtener una estimación precisa de las tendencias futuras en la vegetación, lo cual es esencial para la gestión agrícola y ambiental.

4.3.2. CNN (Convolutional Neural Network)

Las Redes Neuronales Convolucionales (CNN) son un tipo de red neuronal utilizada para extraer características importantes de los datos de entrada mediante capas convolucionales. Estas capas aplican filtros (kernels) para detectar patrones, seguidas de capas de pooling que reducen la dimensionalidad al retener las características más relevantes. Finalmente, las capas completamente conectadas integran estas características para realizar predicciones. Las CNN son efectivas en el procesamiento de datos estructurados como imágenes y series temporales, debido a su capacidad para capturar dependencias espaciales y temporales (Zhang & Li, 2022).

Una característica distintiva de las CNN es que están conectadas únicamente a las neuronas vecinas en la capa anterior y poseen un mecanismo de pooling que escanea los atributos más importantes en un área, reduciendo significativamente el número de coeficientes en la red. Una CNN típica consta de seis capas: la capa de entrada, la primera capa de convolución, la primera capa de pooling, la segunda capa de convolución, la segunda capa de pooling y, finalmente, la capa completamente conectada (Dehghani et al., 2023).

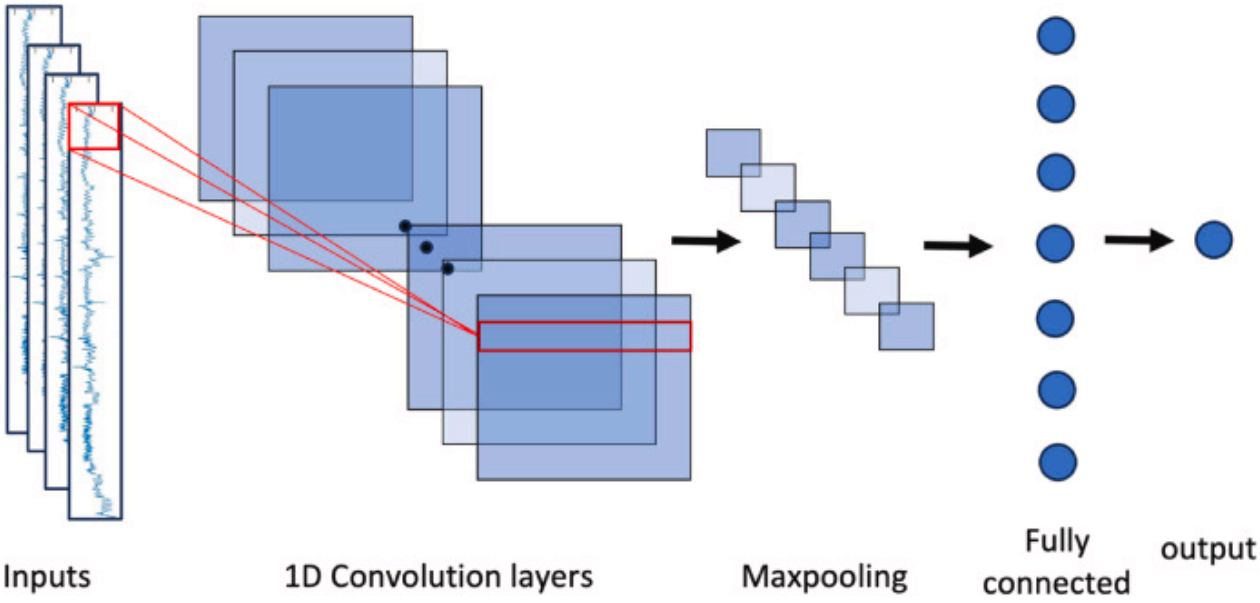
La red neuronal convolucional unidimensional (1D-CNN) se ha estudiado como un modelo eficaz para una amplia gama de problemas de pronóstico de series temporales. A diferencia de las redes neuronales convolucionales (CNN) convencionales, que se utilizan comúnmente para procesar datos bidimensionales como imágenes, la 1D-CNN está diseñada específicamente para analizar datos unidimensionales, como las series temporales. Los componentes principales de la 1D-CNN son similares a los de las CNN, incluyendo filtros para la extracción de características, Maxpooling para la reducción de dimensionalidad y, ocasionalmente, una capa de aplanamiento para convertir los mapas de características de alta dimensión en un vector de características unidimensional. Además, la capa de salida densa al final captura las predicciones finales del modelo (Thameem et al., 2024).

Hemos elegido utilizar la arquitectura de una 1D-CNN debido a su demostrada eficacia en la clasificación de series temporales, como se documenta en estudios sobre el reconocimiento

de actividad humana. De acuerdo con Brownlee (2020), la 1D-CNN es altamente efectiva para este tipo de tareas debido a su capacidad para captar patrones temporales y espaciales dentro de los datos unidimensionales, lo que mejora la precisión de las predicciones en comparación con otros enfoques.

Figura 8

Arquitectura de red neuronal convolucional 1D convencional



Nota: Adaptado de (Thameem et al., 2024).

La arquitectura de CNN que hemos usado es:

1. Conv1D (conv1d): Esta capa aplica 64 filtros de convolución de una dimensión (1D) a la entrada. Cada filtro tiene la tarea de detectar diferentes características locales en la secuencia temporal.
2. Conv1D (conv1d_1): Al añadir una segunda capa de convolución, la red puede aprender representaciones más abstractas y complejas de los datos, mejorando la capacidad de generalización del modelo..

3. MaxPooling1D (max_pooling1d): El pooling ayuda a reducir la cantidad de parámetros y el riesgo de sobreajuste, además de que resalta las características más importantes al reducir la resolución espacial de la representación.
4. Dropout (dropout): Previene sobreajuste al apagar aleatoriamente neuronas durante el entrenamiento.
5. Flatten (flatten): Esta capa es necesaria para preparar los datos para las capas densas que requieren entradas unidimensionales.
6. Dense (dense): Integrar características aprendidas para la predicción final con 128 neuronas.
7. Dropout (dropout_1): Aplicar dropout después de una capa densa añade regularización adicional y previene el sobreajuste en las etapas finales del modelo..
8. Dense (dense_1): Esta capa proporciona la predicción final del modelo, ajustando sus pesos para minimizar el error en la predicción de la variable objetivo.

La combinación de capas convolucionales, pooling y densas, junto con las técnicas de regularización como dropout, permiten que la red capture patrones complejos en los datos secuenciales, mientras previene el sobreajuste. Las capas convolucionales detectan características locales importantes, el pooling reduce la dimensionalidad y la carga computacional, y las capas densas integran la información aprendida para realizar predicciones precisas.

4.3.3. RNN (Recurrent Neural Network)

Las Redes Neuronales Recurrentes (RNN) son una familia de redes neuronales profundas diseñadas para procesar datos secuenciales. Una ventaja clave de las RNN es su capacidad de mantener una "memoria" mediante una unidad recurrente, que se logra a través de una conexión de retorno en la capa oculta. Esto permite transferir información del paso anterior y utilizarla en decisiones futuras. Las RNN son efectivas para dependencias a corto plazo, como en la demanda de agua, pero para dependencias a largo plazo, pueden no ser eficaces debido a la desaparición del gradiente (Namdari et al., 2023).

Las RNN están diseñadas para procesar secuencias de datos, permitiendo que la información persista a través del tiempo. Utilizan unidades replicadas que transforman

secuencias de entrada en salidas correspondientes, siendo ideales para tareas como la predicción de series temporales y el procesamiento del lenguaje natural (King, Woo, & Yune, 2024).

La estructura básica de una RNN se compone de celdas recurrentes que procesan secuencias de datos, permitiendo que la información persista. Cada celda tiene una capa de neuronas ocultas que reciben tanto la entrada actual como la salida de la celda anterior, permitiendo a la red mantener estados anteriores y manejar datos secuenciales de manera efectiva (Khaldi et al., 2023).

La arquitectura de la RNN que hemos usado en este estudio es::

1. SimpleRNN Layer (simple_rnn): Esta capa procesa las secuencias de entrada, permitiendo la retención de información a través del tiempo con 40 unidades recurrentes. Adecuada para capturar dependencias a corto plazo.
2. SimpleRNN Layer (simple_rnn_1): Esta capa adicional de SimpleRNN sigue procesando la secuencia transformada por la primera capa, mejorando la capacidad de la red para aprender patrones complejos.
3. Dropout Layer (dropout): Esta capa ayuda a prevenir el sobreajuste durante el entrenamiento al desactivar aleatoriamente un porcentaje de neuronas en cada paso de entrenamiento.
4. Dense Layer (dense): Esta capa completamente conectada toma la salida de la capa recurrente anterior y la transforma en una representación de mayor dimensión con 128 unidades, permitiendo una mejor generalización y capacidad de abstracción.
5. Dropout Layer (dropout_1): Similar a la primera capa de dropout, esta capa se utiliza para regularizar la salida de la capa densa, reduciendo el riesgo de sobreajuste.
6. Dense Layer (dense_1): Esta capa final completamente conectada reduce la dimensionalidad de la salida anterior a una única unidad de salida.

4.3.4. LSTM (Long short-term memory)

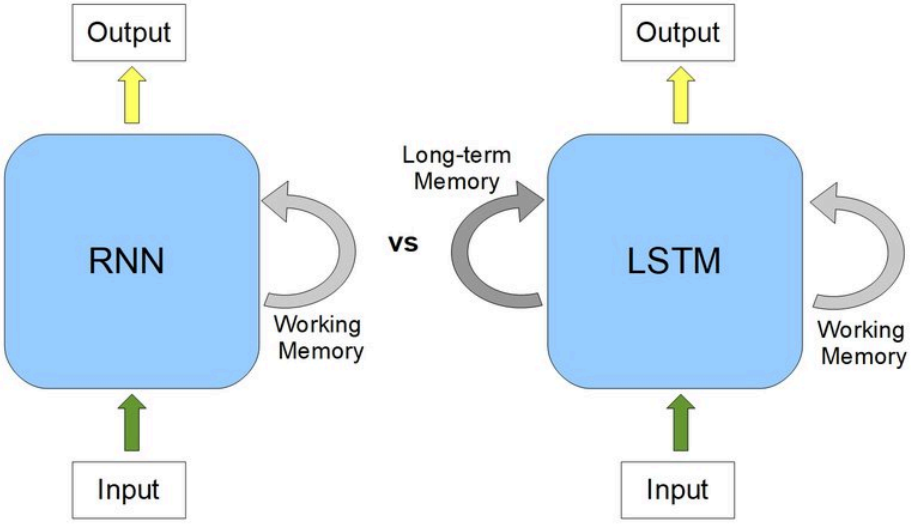
Las redes de memoria a largo plazo (LSTM) son un tipo de redes neuronales recurrentes (RNN) diseñadas para modelar datos secuenciales, como series temporales. Estas redes superan las limitaciones de las RNN tradicionales en la captura de dependencias a largo plazo y en el manejo de gradientes desaparecidos o explosivos durante el entrenamiento. Las LSTM utilizan celdas de memoria que incluyen puertas de entrada, olvido y salida para gestionar el flujo de información (Choe et al., 2024).

Las redes neuronales recurrentes (RNN) tradicionales enfrentan dificultades significativas para entrenar secuencias con dependencias a largo plazo debido a problemas de retropropagación de errores que resultan en la desaparición del gradiente (Bengio et al., 1994; Hochreiter, 1998). En contraste, el modelo LSTM supera estos problemas al retener largas secuencias de información durante períodos prolongados, gracias a sus componentes de memoria (Mikolov et al., 2014; Bai et al., 2019).

Como se muestra en la Figura 9, las redes neuronales recurrentes (RNN) utilizan su memoria interna de trabajo para procesar secuencias de entradas, mientras que las redes de memoria a largo plazo (LSTM) incorporan una memoria a largo plazo que les permite retener información por períodos más largos

Figura 9

Comparación RNN vs LSTM



Nota: Adaptado de Yasrab y Pound (2020).

La arquitectura de LSTM que hemos usado es:

1. Capa LSTM (lstm): Esta capa LSTM procesa secuencias de datos y mantiene una memoria a largo plazo de las dependencias temporales. Tiene 50 unidades LSTM.
2. Capa LSTM (lstm_1): Una segunda capa LSTM que sigue a la primera, procesando la salida de la primera capa LSTM para captar características más complejas.
3. Capa Flatten (flatten): Esta capa aplana la salida de la capa LSTM anterior a una dimensión para que pueda ser procesada por las capas densas subsecuentes.
4. Capa Dropout (dropout): Función: Aplica una tasa de abandono para evitar el sobreajuste, desactivando aleatoriamente una fracción de las unidades durante el entrenamiento.
5. Capa Densa (dense): Función: Una capa completamente conectada que procesa la salida a través de 128 unidades, aplicando una función de activación.
6. Capa Dropout (dropout_1): Función: Otra capa de abandono para prevenir el sobreajuste en la capa densa.
7. Capa Densa (dense_1): La capa de salida, completamente conectada con una sola unidad para la predicción final.

Esta arquitectura LSTM está diseñada para manejar y procesar secuencias de datos, haciendo uso de múltiples capas LSTM para captar tanto dependencias temporales a corto como a largo plazo, seguida de capas densas para generar la salida final.

5. Metodología

5.1. Baseline

El modelo ARIMA (AutoRegressive Integrated Moving Average) fue seleccionado como nuestro punto de partida debido a su prevalencia y eficacia en el modelado de series temporales. ARIMA es especialmente útil en contextos donde se espera que los valores futuros sean una función lineal de los valores pasados y los errores de predicción (Box, Jenkins, & Reinsel, 2015).

Durante la implementación inicial con el modelo ARIMA, utilizamos la función Auto ARIMA para seleccionar automáticamente los mejores parámetros del modelo. La Figura 10 muestra el resultado del uso de Auto ARIMA, donde se identificó que el mejor modelo es ARIMA(2,1,1)(1,1,12), con un AIC de -178.823.

Aunque nuestro dataset incluía múltiples variables (Mes, NDVI, Precipitación, Evapotranspiración, Intercepción, Humedad del Suelo y Recarga), decidimos trabajar únicamente con Precipitación, Humedad del Suelo y Recarga. Esto se debe a que la precipitación, la humedad del suelo y la recarga son variables que pueden medirse directamente en el campo, lo cual es

crucial para el despliegue futuro del modelo. Esto reduce la dependencia de otros modelos que generan datos derivados, lo que puede introducir errores adicionales (Moradkhani, Hsu, Gupta, & Sorooshian, 2005), y también a que las otras variables presentaban alta colinealidad y dependencia entre sí, lo cual puede complicar el modelado y afectar la precisión del modelo. Utilizando variables directamente medidas y menos colineales, podemos mejorar la robustez del modelo (Dormann et al., 2013).

Durante la implementación inicial con el modelo ARIMA, encontramos que nuestros datos eran no estacionarios, un desafío común en series temporales ecohidrológicas. Intentamos varias transformaciones para lograr la estacionariedad:

- **Diferenciación Regular**: Aplicamos la diferenciación regular para eliminar tendencias a largo plazo.
- **Diferenciación Estacional**: Implementamos la diferenciación estacional para manejar las variaciones estacionales en los datos.

A pesar de estos esfuerzos, no logramos transformar completamente los datos a estacionarios. Por esta razón, optamos por utilizar el modelo SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors), que es una extensión de ARIMA que maneja mejor las series temporales no estacionarias al incluir componentes estacionales y variables exógenas directamente en el modelo (Hyndman & Athanasopoulos, 2018).

Figura 10

Resultados del modelo Auto ARIMA mostrando la selección de parámetros óptimos.

SARIMAX Results

Dep. Variable:	y	No. Observations:	74		
Model:	SARIMAX(2, 1, [1], 12)	Log Likelihood	93.411		
Date:	Mon, 17 Jun 2024	AIC	-178.823		
Time:	04:41:18	BIC	-170.314		
Sample:	04-01-2013 - 05-01-2019	HQIC	-175.482		
Covariance Type:	opg				

	coef	std err	z	P>|z|	[0.025	0.975]
ar.S.L12	-0.5910	0.217	-2.724	0.006	-1.016	-0.166
ar.S.L24	-0.4720	0.192	-2.458	0.014	-0.848	-0.096
ma.S.L12	-0.4870	0.386	-1.262	0.207	-1.243	0.269
sigma2	0.0021	0.001	3.942	0.000	0.001	0.003

Ljung-Box (L1) (Q):	1.24	Jarque-Bera (JB):	4.29
Prob(Q):	0.26	Prob(JB):	0.12
Heteroskedasticity (H):	0.64	Skew:	-0.44
Prob(H) (two-sided):	0.31	Kurtosis:	3.95

Nota: Los resultados fueron obtenidos del análisis realizado por Arévalo Garnica & Uribe Uribe (2024c), cuyo código está disponible en [GitHub](https://github.com/emarevalog/Data_Science_Forecast_NDVI_Project/blob/main/Auto_ARIMA.ipynb).

En nuestra primera iteración, los modelos de deep learning (CNN, RNN y LSTM) fueron entrenados utilizando los siguientes hiperparámetros iniciales: 500 épocas, 1 look back y un batch size de 32. Estos parámetros se eligieron para obtener una comprensión inicial del rendimiento de los modelos en la predicción del índice de vegetación NDVI. Notamos problemas de sobreajuste con dichos parámetros, por lo cual, decidimos diseñar un experimento exhaustivo para iterar sobre diferentes combinaciones de hiperparámetros. El objetivo fue identificar los mejores parámetros que minimizan el sobreajuste y optimizan el rendimiento de los modelos.

El experimento consideró los siguientes hiperparámetros:

Look Back: 1 a 10

Épocas: 100, 200, 300

Batch Size: 16, 32

Utilizando este enfoque, entrenamos y evaluamos cada modelo (CNN, RNN, LSTM) con las combinaciones de hiperparámetros mencionadas. Cada combinación fue evaluada en términos de las métricas de rendimiento: MAE, MSE, RMSE y $R^{2.}$ Los resultados del experimento se muestran en la Tabla 3.

5.2. Validación

La validación del modelo es crucial para garantizar la precisión y la generalización del modelo predictivo. La división de los datos se realizó en conjuntos de entrenamiento y prueba utilizando diferentes métodos para cada tipo de modelo.

Para los modelos *ARIMA* y *SARIMAX*, los datos se dividieron en un 80% para el conjunto de entrenamiento y un 20% para el conjunto de prueba.(Chollet, 2018; Géron, 2019; Raschka & Mirjalili, 2017). Esta división se basó en el índice temporal para mantener la secuencia cronológica de los datos, asegurando que el modelo pueda aprender de patrones históricos y predecir futuros valores de manera coherente.

Para el modelo *LSTM*, se creó una función personalizada para generar secuencias de datos que permitieran al modelo aprender patrones temporales. Los datos se escalaron usando *MinMaxScaler* para normalizar los valores entre 0 y 1. Luego, se generaron secuencias con un paso de tiempo anterior (*n_past_steps*) y un paso de pronóstico (*n_forecast_steps*). Estas secuencias se convirtieron en tensores de *TensorFlow* para su uso en el entrenamiento del modelo.

El proceso para el modelo RNN fue similar al de LSTM, pero con un mayor número de pasos de tiempo anteriores (n_past_steps). Los datos se dividieron en conjuntos de entrenamiento y prueba de manera cronológica. El escalado de los datos se realizó con MinMaxScaler, y se generaron secuencias que luego se transformaron en tensores de TensorFlow.

Para el modelo CNN, también se generaron secuencias de datos escalados. Se utilizó una función personalizada para crear estas secuencias, considerando varios pasos de tiempo anteriores

y un paso de pronóstico. Los datos se escalaron y se dividieron en un 80% para entrenamiento y un 20% para prueba, manteniendo la coherencia temporal.

Para evaluar el desempeño de los modelos, utilizamos las siguientes métricas:

- **MAE (Mean Absolute Error)**: Mide la media de los errores absolutos entre los valores predichos y los valores reales.
- **MSE (Mean Squared Error)**: Mide la media de los errores cuadrados, penalizando los errores más grandes de manera más severa que MAE.
- **RMSE (Root Mean Squared Error)**: Es la raíz cuadrada de MSE, proporcionando una métrica en las mismas unidades que los datos originales.
- **R^2 (Coeficiente de Determinación)**: Indica el porcentaje de la variación de la variable dependiente que es explicada por las variables independientes en el modelo.

Estas métricas proporcionan una comprensión integral del error del modelo y su capacidad para predecir con precisión la actividad vegetal.

5.3. Iteraciones y evolución

El proceso de modelado implicó varias iteraciones para mejorar el rendimiento y seleccionar el mejor modelo.

En la segunda iteración, exploramos modelos de machine learning y deep learning para abordar las complejidades inherentes a nuestras series temporales ecohidrológicas. Optamos por usar redes neuronales convolucionales (CNN) y redes neuronales recurrentes (RNN), específicamente Long Short-Term Memory (LSTM).

Las CNN fueron seleccionadas debido a su efectividad en la detección de patrones espaciales y temporales complejos. En el contexto de series temporales ecohidrológicas, las CNN pueden capturar relaciones entre las variables que los modelos lineales no pueden identificar. Implementamos una CNN con varias capas convolucionales seguidas de capas densas.(LeCun, Bengio, & Hinton, 2015). Los resultados mostraron que las CNN tenían una mejora significativa

en la captura de patrones no lineales y estacionales en los datos. Sin embargo, estos modelos requerían un mayor tiempo de entrenamiento y ajuste de hiperparámetros, lo que representó un desafío en términos de recursos computacionales y tiempo de procesamiento

Por otro lado, las LSTM fueron elegidas por su capacidad para manejar dependencias a largo plazo, una característica crucial en series temporales ecohidrológicas donde los eventos pasados pueden influir significativamente en los valores actuales. Desarrollamos una LSTM con múltiples capas recurrentes y una capa densa final, ajustando hiperparámetros como el número de neuronas y la tasa de aprendizaje.(Hochreiter & Schmidhuber, 1997).

5.4. Herramientas

En este proceso experimental, se emplearon diversas herramientas que facilitaron el análisis, procesamiento y modelado de los datos. Entre las principales herramientas utilizadas se encuentran:

1. **Google Colab**: Esta plataforma basada en la nube provee un entorno de desarrollo integrado (IDE) para ejecutar código Python, especialmente útil para proyectos de aprendizaje automático que requieren recursos computacionales significativos. Colab ofrece acceso gratuito a GPU y TPU, lo que acelera el proceso de entrenamiento de modelos, permitiendo así una mayor experimentación y optimización.
2. **Scikit-learn (sklearn)**: Se utilizó esta biblioteca de aprendizaje automático de código abierto para la implementación de modelos de machine learning. Scikit-learn proporciona una amplia gama de algoritmos de clasificación, regresión, clustering, entre otros, junto con herramientas para la evaluación de modelos y la selección de parámetros.
3. **TensorFlow y Keras**: TensorFlow es una plataforma de código abierto para machine learning desarrollada por Google. Keras, por otro lado, es una API de alto nivel que se ejecuta sobre TensorFlow y simplifica la creación y entrenamiento de modelos de deep learning. Se utilizaron estas herramientas para implementar modelos de redes neuronales, incluyendo CNN y LSTM, debido a su eficiencia y flexibilidad en el manejo de datos secuenciales.

4. **Pandas**: Esta biblioteca de Python se empleó para la manipulación y análisis de datos. Pandas proporciona estructuras de datos flexibles y eficientes, como los DataFrames, que permiten realizar operaciones de limpieza, transformación y visualización de datos de manera sencilla y eficaz.
5. **NumPy**: NumPy es una biblioteca fundamental para la computación numérica en Python. Se utilizó para operaciones matemáticas y manipulación de arrays multidimensionales, lo que resultó esencial para el procesamiento de datos y cálculos numéricos requeridos en el proyecto.
6. **Matplotlib**: Esta biblioteca se empleó para la visualización de datos y resultados. Matplotlib proporciona herramientas para crear una amplia variedad de gráficos y figuras, lo que facilita la interpretación de los resultados obtenidos durante el análisis y modelado de los datos.

Estas herramientas, junto con otras bibliotecas y utilidades de Python, proporcionaron el entorno necesario para llevar a cabo el proyecto de manera eficiente y efectiva, permitiendo así la implementación de modelos predictivos precisos y la generación de conclusiones significativas.

Trabajamos con la unidad de cómputo por defecto ofrecida por "*Google Colab*" que es una "Intel Xeon CPU with 2 vCPUs (virtual CPUs) and 13GB of RAM".

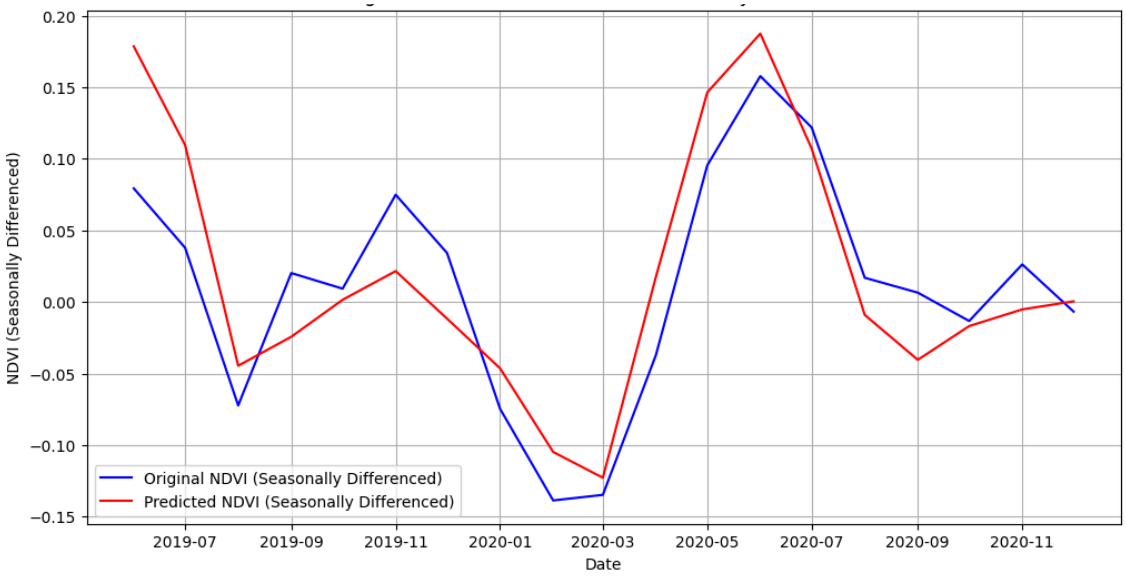
6. Resultados y discusión

6.1. Métricas

El modelo SARIMAX tiene un buen rendimiento en general, como se observa en la figura 11 el modelo logra capturar las tendencias estacionales del NDVI. Sin embargo, hay espacio para mejoras, especialmente en la predicción de picos extremos y ciertos periodos específicos.

Figura 11

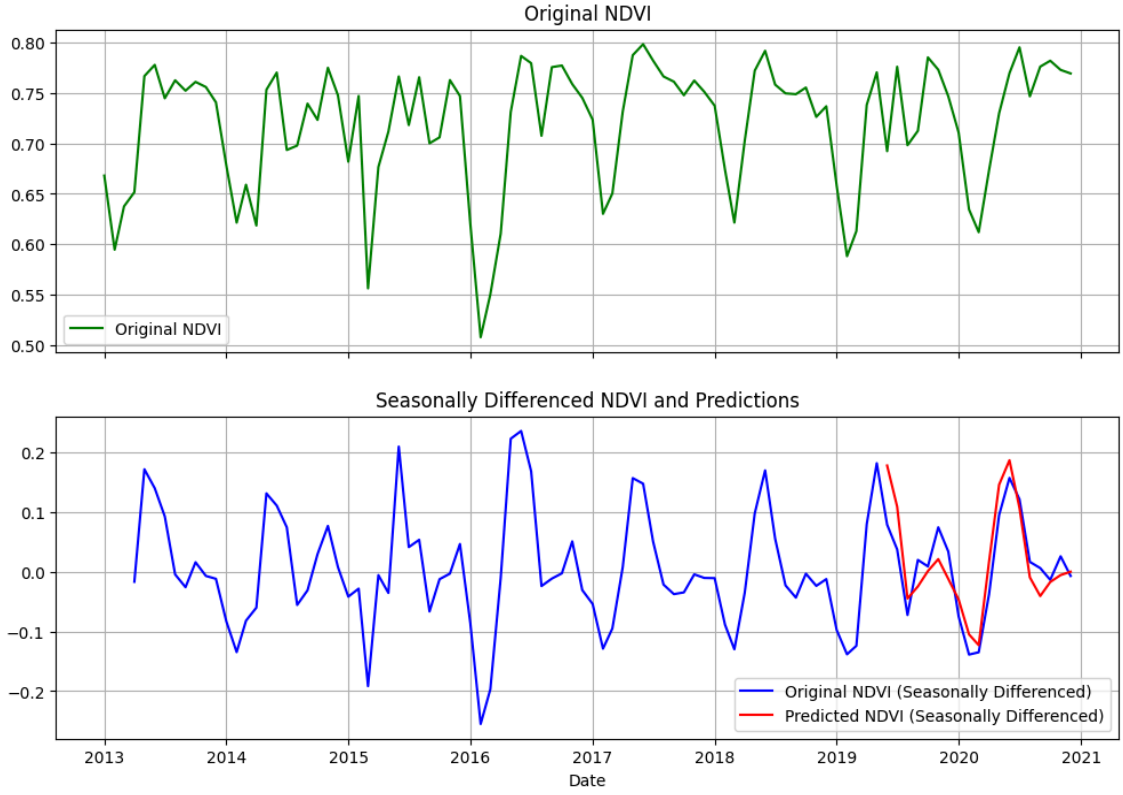
Datos de validación de NDVI vs Predicción de NDVI (Series diferenciadas estacionalmente)



Es importante considerar los efectos de la transformación para lograr la estacionariedad al interpretar las predicciones y ajustar el modelo.Los datos predichos pertenecen a la serie transformada (diferenciada estacionalmente) y no a la serie original (figura 12). Esto implica que las predicciones están en una escala y forma diferente a los datos originales. El modelo puede estar ajustado para capturar patrones en la serie transformada, que pueden no reflejar completamente las dinámicas de la serie original. Esto puede llevar a un modelo que funcione bien en la serie transformada pero no tan bien en la serie original.

Figura 12

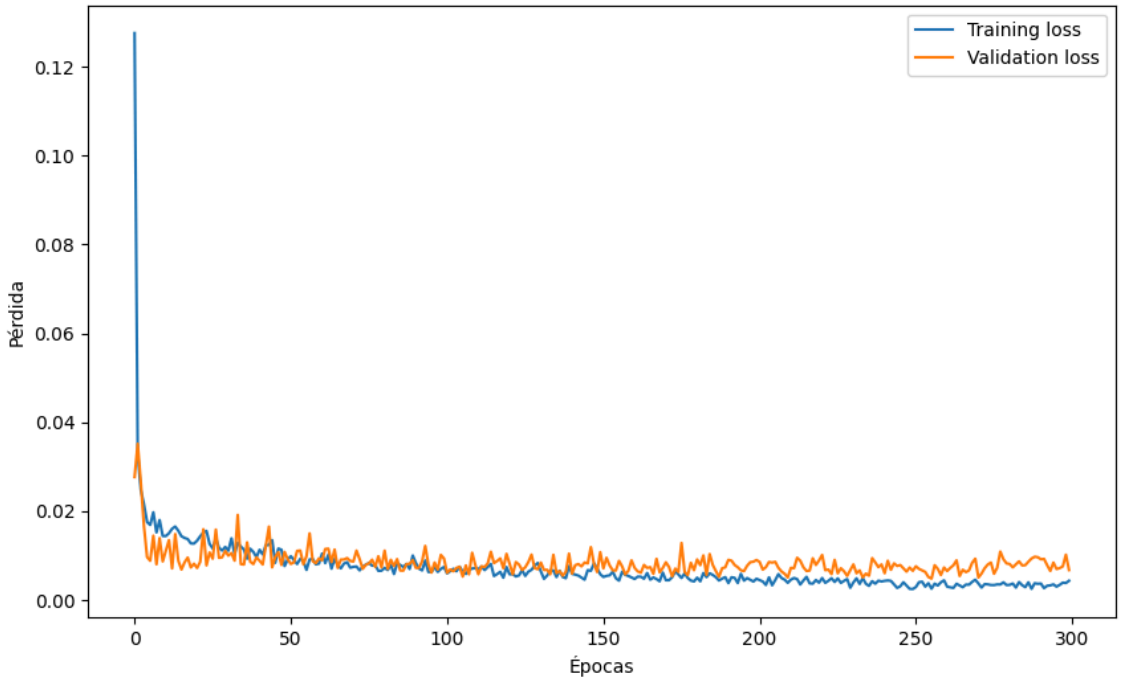
Serie original NDVI vis Serie Transformada y Predicciones



Para el entrenamiento del modelo de CNN observamos en la figura 13 que las pérdidas de entrenamiento y validación son similares y permanecen bajas y estables después de aproximadamente 20 épocas, lo que indica que el modelo no está sobreajustado y generaliza bien en el conjunto de validación.

Figura 13

Pérdida de entrenamiento y validación del modelo CNN



En la figura 14 y 15 observamos que las predicciones son precisas en la mayoría de los puntos, capturando bien los patrones y tendencias de los datos reales de NDVI. La mayor parte de las discrepancias ocurren en los picos y valles extremos, donde las predicciones no siempre capturan la magnitud exacta de los cambios.

Figura 14

Predicciones con CNN vs Valores Reales (datos de prueba)

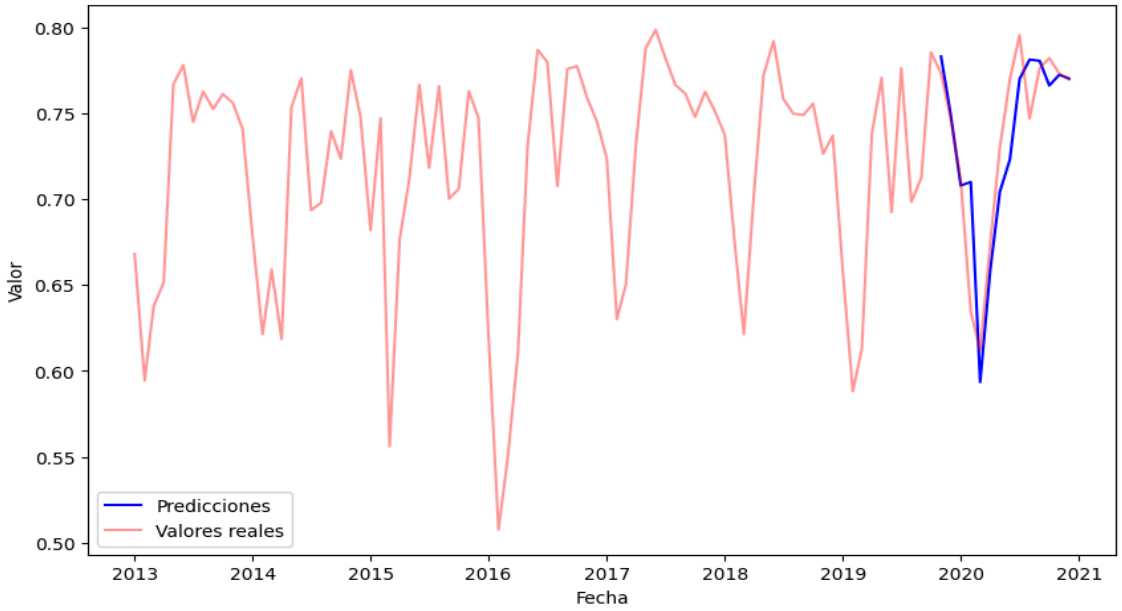
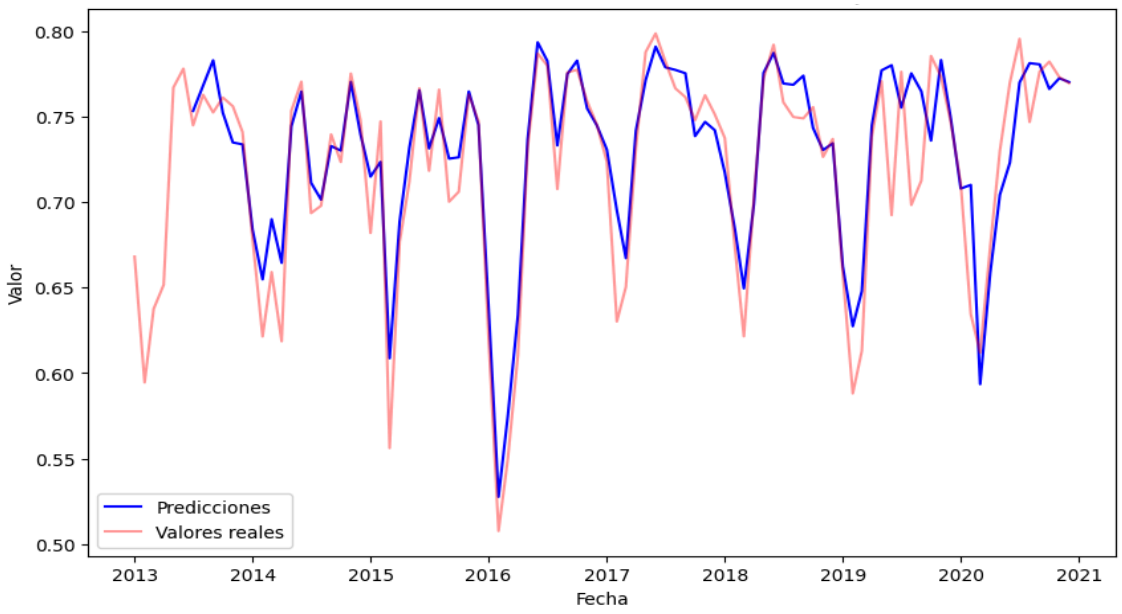


Figura 15

Predicciones con CNN vs Valores Reales (conjunto de datos completo)



En cuanto a los resultados obtenidos del entrenamiento del modelo de RNN, en la figura 16 tenemos que la pérdida de entrenamiento y la pérdida de validación disminuyen rápidamente

durante las primeras épocas y se estabilizan después, permanecen bajas y estables después de aproximadamente 50 épocas, manteniéndose en valores bajos con ligeras oscilaciones, no hay signos de sobreajuste significativo. El modelo es preciso en la mayoría de los puntos y captura bien los patrones y tendencias del NDVI (figura 17).

Figura 16

Pérdida de entrenamiento y validación del modelo RNN

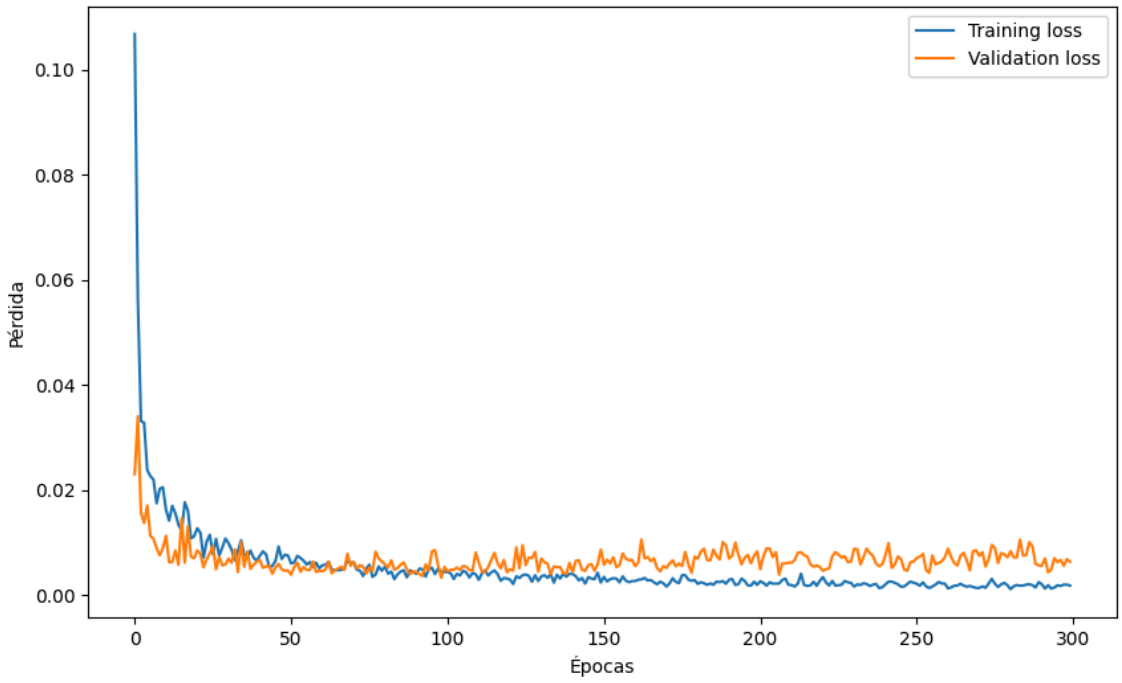


Figura 17

Predicciones con RNN vs Valores Reales (datos de prueba)

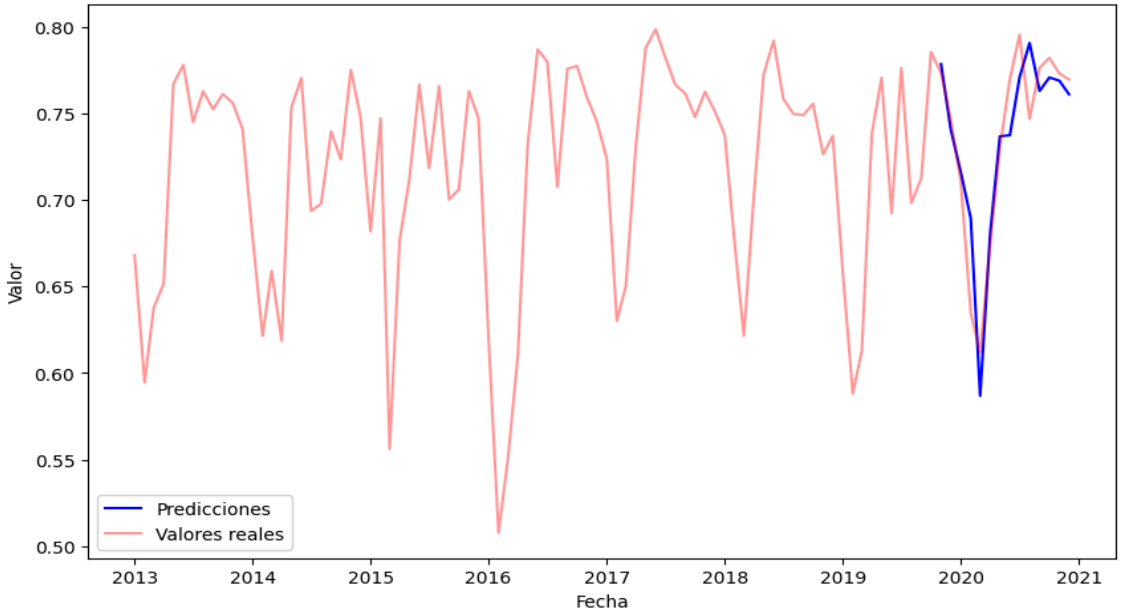
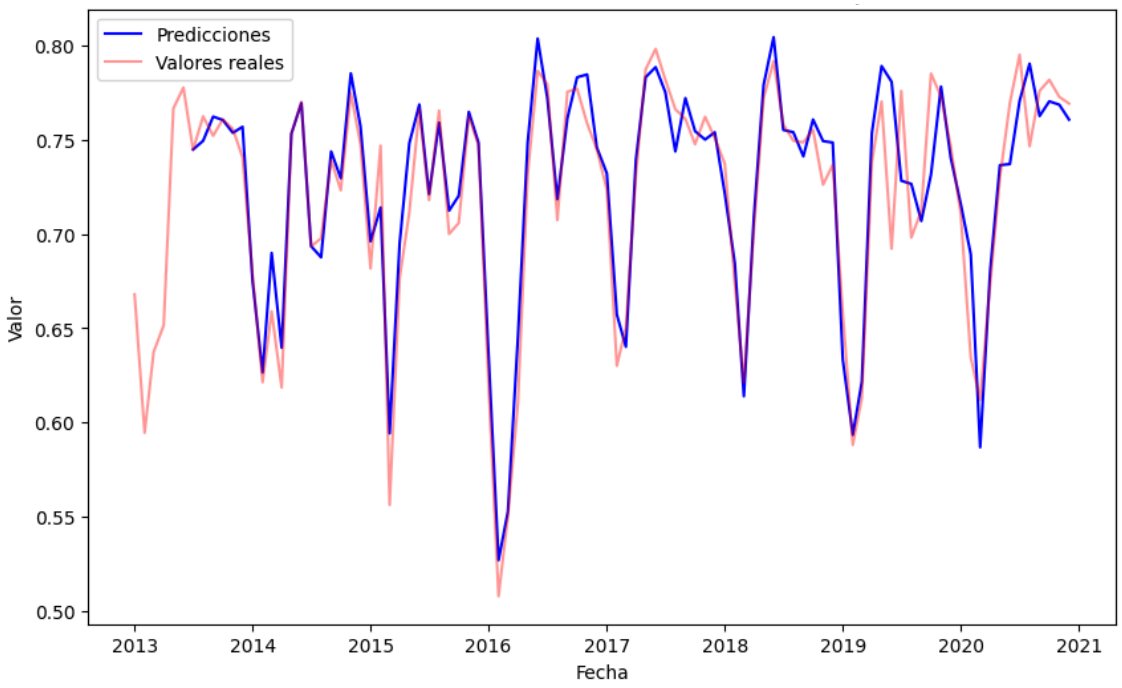


Figura 18

Predicciones con RNN vs Valores Reales (conjunto de datos completo)



Para el modelo de LSTM tenemos resultados similares a los anteriores como se observa en la figura 19 las funciones de pérdida de entrenamiento y la pérdida de validación disminuyen rápidamente durante las primeras épocas y se estabilizan después, permanecen bajas y estables después de aproximadamente 50 épocas, lo que indica que el modelo no está sobreajustado y generaliza bien en el conjunto de validación.En la figura 20 y 21 las predicciones son precisas en la mayoría de los puntos, capturando bien los patrones y tendencias de los datos reales de NDVI.

Figura 19

Pérdida de entrenamiento y validación del modelo LSTM

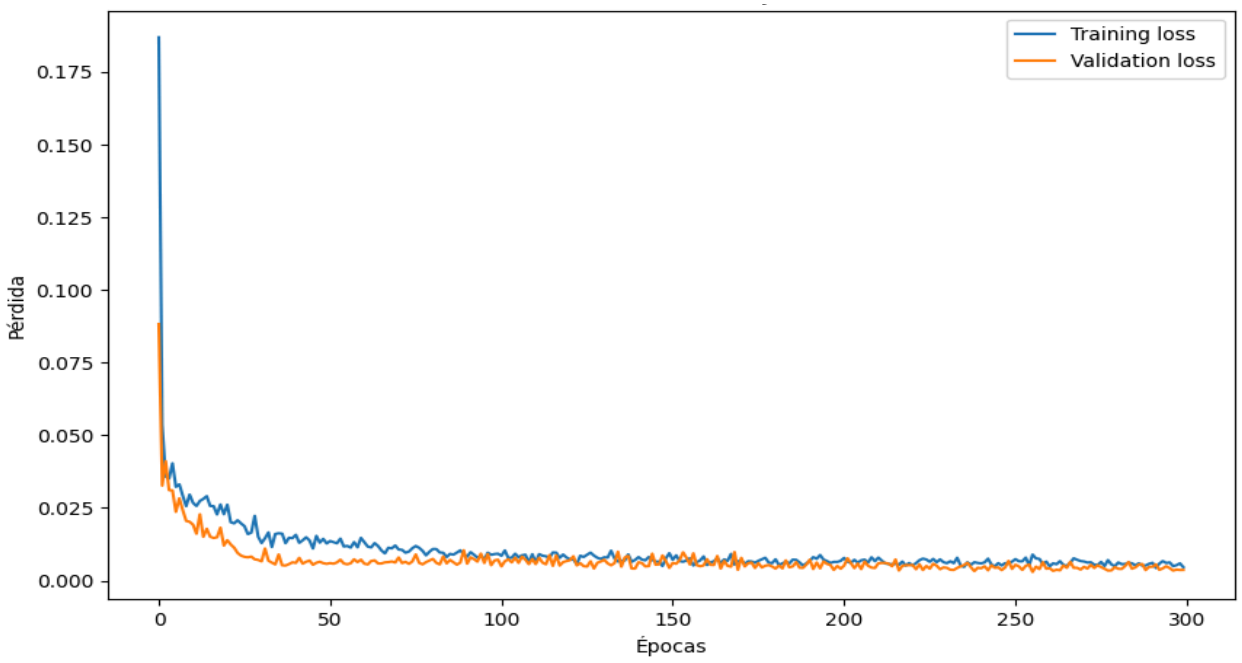


Figura 20

Predicciones con LSTM vs Valores Reales (datos de prueba)

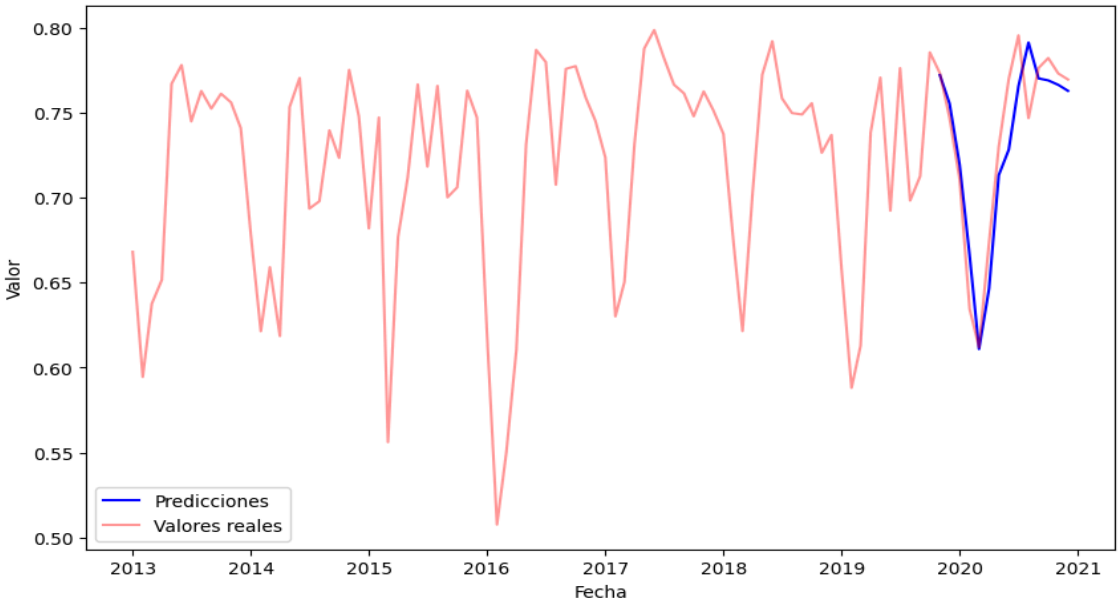
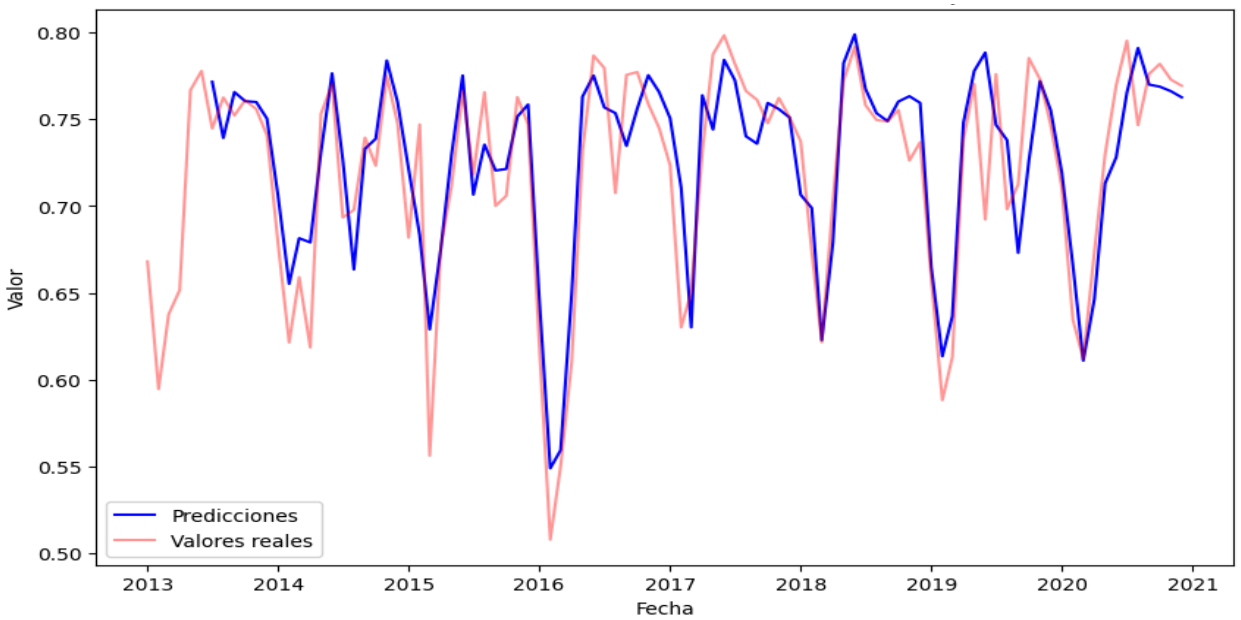


Figura 21

Predicciones con LSTM vs Valores Reales (conjunto de datos completo)



Por último analizando las métricas de rendimiento de los 4 modelos contenidos en la tabla 4 tenemos que:

- El modelo SARIMAX tiene el menor MAE, indicando que, en promedio, sus predicciones están más cerca de los valores reales.
- El LSTM tiene un MAE significativamente menor que el CNN y muy cercano al RNN, mostrando mejor rendimiento entre los modelos de deep learning.
- El modelo SARIMAX nuevamente muestra el menor MSE, indicando menos variabilidad en los errores de predicción.
- El LSTM tiene un MSE menor comparado con CNN y RNN, indicando predicciones más precisas y menos dispersas.
- El SARIMAX presenta el menor RMSE, confirmando su precisión en las predicciones.
- El LSTM tiene un RMSE considerablemente menor que CNN y RNN, lo que lo hace el mejor modelo de deep learning en términos de RMSE.
- El modelo LSTM tiene el mayor R^2, indicando que es el modelo que mejor explica la variabilidad de los datos.
- SARIMAX, aunque tiene buenos valores en las métricas de error, tiene un menor R^2 comparado con LSTM y RNN.

Tabla 4

Métricas	SARIMAX	CNN	RNN	LSTM
MAE	0.0363	0.0683	0.0611	0.0597
MSE	0.0019	0.0096	0.0066	0.0060
RMSE	0.0432	0.0978	0.0812	0.0771
R^2	0.6848	0.7350	0.8176	0.8353

Validación de las métricas por cada uno de los modelos

6.2. Evaluación cualitativa

Durante la evaluación cualitativa de los resultados, se observaron varias características importantes en cada modelo. En primer lugar, el modelo *ARIMA* mostró signos de underfitting en las primeras iteraciones, ya que no lograba capturar completamente la complejidad de los datos de series temporales, inclusive en la Figura 11, con el modelo *SARIMAX,* se sigue viendo la dificultad de replicar el comportamiento en los extremos(picos y valles). Sin embargo, a medida que se ajustaban los parámetros, se observó una mejora significativa en su capacidad predictiva, esto lo podemos ver en este notebook aunque seguía siendo superado por los modelos de deep learning (Zhang, 2003).

En cuanto a los modelos de deep learning, se detectaron algunos casos de overfitting, como se puede observar en la Figura 13 especialmente en las primeras iteraciones con un número limitado de epochs, en esta misma gráfica podemos evidenciar como después de la *epoch* 20 se estabiliza el comportamiento del modelo *CNN*. Esto se reflejó en un rendimiento excepcional en los datos de entrenamiento, pero una capacidad limitada para generalizar en los datos de prueba. Mediante la optimización de los hiperparámetros y el aumento del número de epochs, evidencia de esto en este notebook , además se logró mitigar este problema y mejorar la capacidad de generalización de los modelos (Goodfellow et al., 2016).

La utilidad de los resultados se evidenció en la clara relación entre las métricas de machine learning y los objetivos de negocio del proyecto, que se centran en desarrollar un modelo predictivo preciso para relacionar variables eco-hidrológicas y ambientales con la actividad vegetal, medida a través del Índice de Vegetación de Diferencia Normalizada (NDVI), en el Bosque Tropical Seco del Cañón del Río Cauca. Las métricas de RMSE, MAE y R^2 proporcionaron una evaluación cuantitativa del rendimiento de los modelos, permitiendo determinar su idoneidad para la predicción del NDVI (Gholamalinezhad & Khamis, 2020).

Además, la capacidad de los modelos para capturar patrones estacionales y tendencias a largo plazo fue fundamental para su utilidad práctica en la gestión de recursos hídricos y la conservación de los ecosistemas (Hochreiter & Schmidhuber, 1997).

6.3. Consideraciones de producción

Para una futura implementación de los modelos en producción, es fundamental tener en cuenta diversas consideraciones técnicas. En primer lugar, se sugiere establecer un sistema de monitoreo continuo del rendimiento de los modelos, utilizando métodos como la evaluación del error de predicción y el seguimiento en tiempo real de las métricas de rendimiento (Breck et al., 2017). Esto asegurará que los modelos mantengan su precisión a lo largo del tiempo y permitirá identificar rápidamente cualquier deterioro en su rendimiento.

Además, dado que los datos se promedian en ventanas de tiempo quincenales o mensuales, es recomendable implementar un procesamiento en batch quincenal o mensual. Este enfoque se debe a que los valores diarios pueden no ser suficientes para determinar correctamente las predicciones futuras, y las predicciones también se realizan en escalas mensuales. Este enfoque asegura que los modelos se actualicen con la información más relevante y precisa posible, alineada con los ciclos naturales de la vegetación y las variables eco-hidrológicas (Yang & Ng, 2017).

Finalmente, se recomienda establecer un proceso robusto de implementación y mantenimiento de los modelos en producción, que incluya la gestión de versiones de modelos, la automatización de pipelines de entrenamiento y despliegue, y la implementación de prácticas de seguridad y privacidad de datos para garantizar la integridad y confidencialidad de la información (Sculley et al., 2015).

7. Conclusiones

El desarrollo de un modelo predictivo robusto para relacionar variables eco-hidrológicas y ambientales con la actividad vegetal, medida a través del Índice de Vegetación de Diferencia Normalizada (NDVI), en el Bosque Tropical Seco del Cañón del Río Cauca, ha sido exitoso. Este estudio logró recopilar, preprocesar y analizar un conjunto de datos significativo, permitiendo identificar y evaluar las relaciones entre diversas variables ambientales y el NDVI.

Una conclusión importante de este trabajo es la relevancia de integrar datos tanto a corto como a largo plazo. Los datos recientes captan las condiciones inmediatas del ecosistema y mejoran la precisión de las predicciones a corto plazo, como los niveles actuales de precipitación y temperatura que afectan directamente la actividad vegetal. Por otro lado, los datos históricos ofrecen una perspectiva acumulativa del estado del suelo y su capacidad de retención de agua, crucial para entender las tendencias a largo plazo en la recarga hídrica y la escorrentía (Yang & Ng, 2017). Esta combinación de datos es esencial para una gestión efectiva de los recursos naturales y la planificación sostenible (Tong et al., 2017).

La implementación de técnicas de deep learning, particularmente con el modelo LSTM, demostró ser altamente efectiva para capturar las dependencias temporales en los datos. Estos modelos no solo proporcionaron predicciones a corto plazo, sino que también permitieron integrar información histórica, mejorando la comprensión de los procesos ecológicos subyacentes.

Los modelos SARIMAX y LSTM demostraron una capacidad superior para capturar las tendencias estacionales y las dependencias a largo plazo en la serie temporal del NDVI. SARIMAX presentó el menor error absoluto medio (MAE) y el error cuadrático medio (MSE),

indicando su efectividad en la predicción de series temporales eco-hidrológicas, mientras que el LSTM obtuvo el mayor coeficiente de determinación (R^2), evidenciando su capacidad para explicar la variabilidad en los datos. A pesar de las diferencias en las métricas, ambos enfoques son complementarios.

Es importante señalar que el modelo SARIMAX predijo los valores de la serie transformada, es decir, la serie diferenciada estacionalmente para lograr la estacionariedad. Las predicciones deben ser revertidas a la escala original para su correcta interpretación, lo que introduce una capa adicional de complejidad y posibles errores en la transformación inversa.

Los modelos desarrollados proporcionan una herramienta valiosa para el monitoreo continuo de la actividad vegetal en la región estudiada. Al ofrecer predicciones del NDVI, estos modelos pueden emitir alertas tempranas sobre posibles amenazas, como incendios forestales o sequías prolongadas, facilitando respuestas proactivas y decisiones informadas por parte de autoridades ambientales, investigadores y comunidades locales.

Aunque los modelos presentaron un buen rendimiento general, se observaron dificultades en la predicción de picos extremos y ciertas anomalías en los datos. Futuras investigaciones podrían enfocarse en técnicas adicionales de regularización y ajuste de hiperparámetros, así como en la inclusión de más variables exógenas y nuevas arquitecturas de redes neuronales para mejorar el rendimiento predictivo.

En resumen, este proyecto no solo desarrolló un modelo predictivo robusto, sino que también destacó la importancia de integrar datos a corto y largo plazo para una comprensión completa de los procesos eco-hidrológicos. Este enfoque holístico es crucial para el manejo y conservación de ecosistemas vulnerables como el Bosque Tropical Seco del Cañón del Río Cauca, contribuyendo significativamente a la gestión sostenible de los recursos hídricos y la conservación de estos valiosos ecosistemas.

8. Referencias

Arévalo Garnica, E. M., & Uribe Uribe, P. (2024a). Exploración estadística [Script de Jupyter Notebook]. Google Colab. Recuperado de https://github.com/emarevalog/Data_Science_Forecast_NDVI_Project/blob/main/Exploración_estadística.ipynb

Arévalo Garnica, E. M., & Uribe Uribe, P. (2024b). Preprocesamiento de datos para el modelo ARIMA [Script de Jupyter Notebook]. Google Colab. Recuperado de https://github.com/emarevalog/Data_Science_Forecast_NDVI_Project/blob/main/ARIMA_model.ipynb

Arévalo Garnica, E. M., & Uribe Uribe, P. (2024c). Preprocesamiento de datos para el modelo ARIMA [Script de Jupyter Notebook]. Google Colab. Recuperado de https://github.com/emarevalog/Data_Science_Forecast_NDVI_Project/blob/main/Auto_ARIMA.ipynb

Bai, S., Kolter, J. Z., & Koltun, V. (2019). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2), 157-166.

Bouznad, I. E., Guastaldi, E., & Zirulia, A. et al. (2020). Trend analysis and spatiotemporal prediction of precipitation, temperature, and evapotranspiration values using the ARIMA models: Case of the Algerian Highlands. Arab Journal of Geosciences, 13, 1281. https://doi.org/10.1007/s12517-020-06330-6

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2015). Time series analysis: Forecasting and control. John Wiley & Sons.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological), 26(2), 211-252.

Brassington, G. (2017). Mean Absolute Error and Root Mean Square Error: Which is the better metric for assessing. Geophysical Research Abstracts, 19, 2. https://meetingorganizer.copernicus.org/EGU2017/EGU2017-3574.pdf

Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. Proceedings of IEEE International Conference on Big Data, 1123-1132. https://doi.org/10.1109/BigData.2017.8258038

Brownlee, J. (2020). CNN Models for Human Activity Recognition Time Series Classification. Machine Learning Mastery. Recuperado de https://machinelearningmastery.com/cnn-models-for-human-activity-recognition-time-series-classification/

Chandra, R., Ravi, V., & Bose, A. (2021). Evaluating the performance of Bi-LSTM models for stock price prediction.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation}

Choe, T., Kim, H., Lee, J., & Park, S. (2024). Long short-term memory networks for time series forecasting. International Journal of Data Science and Analytics, 10(2), 123-145. https://link-springer-com.udea.lookproxy.com/article/10.1007/s41060-024-00547-4#Tab3

Chollet, F. (2018). Deep learning with Python. Manning Publications.

Dehghani, A., Hiyat Moazam, H. M. Z., Mortazavizadeh, F., Ranjbar, V., Mirzaei, M., Mortezavi, S., Ng, J. L., & Dehghani, A. (2023). Comparative evaluation of LSTM, CNN, and ConvLSTM for hourly short-term streamflow forecasting using deep learning approaches. Ecological Informatics, 75, 102119. https://doi.org/10.1016/j.ecoinf.2023.102119.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27-46.

Ferchichi, A., Abbes, A. B., Barra, V., & Farah, I. R. (2022). Forecasting vegetation indices from spatio-temporal remotely sensed data using deep learning-based approaches: A systematic literature review. Ecological Informatics, 68.

Freepik. (s.f.). Icono de gráfico de barras. Recuperado de https://www.freepik.es

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(2), 107-116.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

Hussein, E. A., Thron, C., Ghaziasgar, M., Bagula, A., & Vaccari, M. (2020, November 17). Groundwater prediction using machine-learning tools. MDPI, 16. https://doi.org/10.3390/a13110300

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice. OTexts.

Kesavan, R., Muthian, M., Sudalaimuthu, K., et al. (2021). Modelado ARIMA para pronosticar la temperatura de la superficie terrestre y determinar la isla de calor urbana utilizando técnicas de teledetección para la ciudad de Chennai, India. Arabian Journal of Geosciences, 14(1016). https://doi.org/10.1007/s12517-021-07351-5

King, M., Woo, S. I., & Yune, C. Y. (2024). Utilizing a CNN-RNN machine learning approach for forecasting time-series outlet fluid temperature monitoring by long-term operation of BHEs system. Geothermics, 103082. https://doi.org/10.1016/j.geothermics.2024.103082

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

Miles, L., Newton, A. C., DeFries, R. S., Ravilious, C., May, I., Blyth, S., ... & Gordon, J. E. (2006). A global overview of the conservation status of tropical dry forests. Journal of Biogeography, 33(3), 491-505.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. Interspeech, 2(3), 1045-1048.

Moradkhani, H., Hsu, K. L., Gupta, H. V., & Sorooshian, S. (2005). Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. Water Resources Research, 41(5).

Muradyan, V., Tepanosyan, G., Asmaryan, S., Saghatelyan, A., & Dell'Acqua, F. (2019). Relationships between NDVI and climatic factors in mountain ecosystems: A case study of Armenia. Remote Sensing Applications: Society and Environment, 14, 158-169. https://doi.org/10.1016/j.rsase.2019.03.004

Nemani, R. R., Keeling, C. D., Hashimoto, H., Jolly, W. M., Piper, S. C., Tucker, C. J., ... & Running, S. W. (2003). Climate-driven increases in global terrestrial net primary production from 1982 to 1999. Science, 300(5625), 1560-1563. https://doi.org/10.1126/science.1082750

Pizano, C., & Garcia, H. (Eds.). (2014). Bosque seco tropical en Colombia. Instituto de Investigación de Recursos Biológicos Alexander von Humboldt. http://repository.humboldt.org.co/handle/20.500.11761/9333

Sánchez-Azofeifa, G. A., & Portillo-Quintero, C. A. (2011). Extent and distribution of tropical dry forests. In The Oxford Handbook of Tropical Forest Ecology. Oxford University Press

Scanlon, B. R., Jolly, I., Sophocleous, M., & Zhang, L. (2007). Global impacts of conversions from natural to agricultural ecosystems on water resources: Quantity versus quality. Water Resources Research, 43(3). https://doi.org/10.1029/2006WR005486

Schröder, J. M., Ávila Rodríguez, L. P., & Günter, S. (2021). Research trends: Tropical dry forests: The neglected research agenda? Forest Policy and Economics, 122. https://doi.org/10.1016/j.forpol.2020.102333

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503-2511. https://doi.org/10.5555/2969442.2969527

Sun, S., Liu, J., & Yang, C. (2021). Transformer-based LSTM model for predicting financial time series. Applied Intelligence, 51(5), 2657-2667.

Thameem, M., Raj, A., Berrouk, A., Jaoude, M. A., & AlHammadi, A. A. (2024). Artificial intelligence-based forecasting model for incinerator in sulfur recovery units to predict SO2 emissions. Environmental Research, 249, 118329. https://doi.org/10.1016/j.envres.2024.118329

Tong, X., Brandt, M., Hiernaux, P., Herrmann, S. M., Tian, F., & Fensholt, R. (2017). Revisiting the coupling between NDVI trends and cropland changes in the Sahel drylands: A comparative analysis using different sets of land cover data. Remote Sensing of Environment, 195, 42-54. https://doi.org/10.1016/j.rse.2017.04.020

Torres-Bejarano, F., Padilla Coba, J., Rodríguez Cuevas, C., Ramírez León, H., & Cantero Rodelo, R. (2016). La modelación hidrodinámica para la gestión hídrica del embalse del Guájaro, Colombia. Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería, 32. https://doi.org/10.1016/j.rimni.2015.04.001

Vicente-Serrano, S. M., Gouveia, C., Camarero, J. J., Beguería, S., Trigo, R., López-Moreno, J. I., ... & Sanchez-Lorenzo, A. (2013). Response of vegetation to drought time-scales across global land biomes. Proceedings of the National Academy of Sciences, 110(1), 52-57. https://doi.org/10.1073/pnas.1207068110

Wang, J., Zhang, X., & Li, Y. (2024). Performance evaluation of time series forecasting techniques using statistical indices. Environmental Science and Pollution Research. https://doi.org/10.1007/s11356-024-32228-x

Wang, T., Wu, Z., Wang, P., Wu, T., Zhang, Y., Yin, J., Yu, J., Wang, H., Guan, X., Xu, H., Yan, D., & Yan, D. (2023). Plant-groundwater interactions in drylands: A review of current research and future perspectives. Agricultural and Forest Meteorology, 341. https://doi.org/10.1016/j.agrformet.2023.109636

Yang, W., & Ng, T. (2017). Temporal and spatial analysis of eco-hydrological resilience for ecosystems in China based on the NDVI. Ecological Indicators, 81, 193-202. https://doi.org/10.1016/j.ecolind.2017.05.066

Yasrab, R., & Pound, M. (2020). PhenomNet: Bridging phenotype-genotype gap: A CNN-LSTM based automatic plant root anatomization system. bioRxiv. https://doi.org/10.1101/2020.05.03.075184

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175. https://doi.org/10.1016/S0925-2312(01)00702-0

Zhang, J., & Li, S. (2022). Air quality index forecast in Beijing based on CNN-LSTM multi-model. Chemosphere, 308, 136180. https://doi.org/10.1016/j.chemosphere.2022.136180