UNIVERSIDAD NACIONAL DE ROSARIO

FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

CARRERA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

Título de la Tesis: Detección de anomalías en cultivos agrícolas en base a información obtenida de sensores remotos. Una comparación entre Redes Neuronales y modelos SARIMA.

Maestrando: Lic. Diaz, Santiago Raul.

Director: Mg. Ing. Tentor, Fernando Raúl.

Co Directora: Mg. Lic. Mendez, Fernanda.

Agradecimientos.

Quisiera expresar mi profundo agradecimiento a la comunidad docente de la Universidad Nacional de Rosario, especialmente a los profesores de la Facultad de Ciencias Económicas y Estadística. Su incansable dedicación y esfuerzo son cruciales para formar profesionales altamente capacitados y contribuir al avance de la ciencia y la tecnología, pilares fundamentales para el desarrollo y bienestar de nuestro país.

Un reconocimiento especial es merecido por el Mg. Fernando Raúl Tentor y la Mg. Fernanda Méndez, mi directora y co-directora de tesis, respectivamente. Su constante apoyo y guía han sido fundamentales en mi camino académico. Gracias a ellos, he superado los retos que implica la realización de mi tesis de Maestría.

El presente logro no solo refleja mi dedicación personal, sino también la calidad y el compromiso de todas las personas que me han acompañado en este camino.

Resumen.

Este trabajo analiza la detección de anomalías en cultivos agrícolas mediante el uso de series temporales de *Normalized Difference Vegetation Index* (NDVI) obtenidas por sensores remotos, comparando el rendimiento de dos modelos predictivos: el modelo estadístico *Seasonal Autoregressive Integrated Moving Average* (SARIMA) y la red neuronal *Long Short-Term Memory* (LSTM). El estudio se aplicó a datos satelitales del sensor Sentinel-2 correspondientes a cultivos de soja en el noreste de la provincia de Santa Fe, Argentina.

Se implementaron técnicas de preprocesamiento como el suavizado con filtros Savitzky-Golay y medias móviles, y se aplicó aumentación artificial de datos para compensar la baja ocurrencia de eventos anómalos. El modelo SARIMA mostró un ajuste adecuado bajo criterios de información como AIC, BIC y HQIC, pero falló en la detección de anomalías abruptas, como la caída del NDVI atribuida a una helada. En contraste, LSTM fue capaz de detectar eventos significativos que superaron el umbral de 3 desviaciones estándar (3σ), demostrando una mayor sensibilidad y adaptabilidad ante patrones no lineales.

Además, se identificaron valores de NDVI mayores a 0.8 que, si bien fueron señalados como atípicos por el modelo, no se interpretaron como errores, sino como eventos inusuales según especialistas agrónomos. Los hallazgos confirman que el modelo LSTM es más adecuado que SARIMA para la detección temprana de anomalías en series temporales NDVI en regiones agrícolas con alta variabilidad climática.

Palabras clave: Series temporales, SARIMA, LSTM detección de anomalías, sensores remotos.

Abstract.

This research focuses on anomaly detection in agricultural crops using NDVI time series data derived from remote sensing, comparing the performance of two predictive models: the **SARIMA** statistical model and the **LSTM** neural network. The study uses Sentinel-2 satellite imagery applied to soybean crops in northeastern Santa Fe, Argentina.

Preprocessing techniques such as Savitzky-Golay smoothing and moving averages were implemented, along with artificial data augmentation to address the scarcity of labeled anomalies. While SARIMA showed reasonable model fit AIC, BIC and HQIC, it failed to detect abrupt changes, such as a NDVI drop due to a frost event. In contrast, LSTM successfully detected events exceeding the **3σ (three standard deviation)** threshold,

demonstrating greater sensitivity to nonlinear patterns.

The model also flagged NDVI values above 0.8, which agronomic experts considered unusual but not erroneous, likely caused by atmospheric effects or sensor noise. Findings support that **LSTM outperforms SARIMA** in detecting anomalies in NDVI time series for agricultural regions with high climatic variability.

Keywords: Time Series, SARIMA, LSTM, Anomaly Detection, Remote Sensing.

Índice de contenidos.

Índice de figuras.

Índice de tablas.

1. Introducción.

La detección de anomalías en series temporales es un campo de gran interés tanto para investigadores como para profesionales, especialmente en el ámbito de la agricultura de precisión. Utilizando imágenes multiespectrales del satélite Sentinel-2 de la Agencia Espacial Europea (ESA), disponibles desde 2015, este estudio se enfoca en el análisis de la vegetación a través del Índice de NDVI, que es esencial para monitorear el crecimiento de los cultivos en tiempo casi real. Según Meroni et al. (2019), el NDVI es crucial para evaluar la salud y el rendimiento de los cultivos, destacando su capacidad para integrar el efecto de variables ambientales que afectan directamente el crecimiento de los cultivos. Los estudios han resaltado las capacidades de las redes neuronales, como las de tipo LSTM, para modelar series temporales capturando correlaciones a largo plazo sin la necesidad de ventanas temporales específicas, lo cual es ventajoso para identificar desviaciones del comportamiento estándar. Sin embargo, persisten dudas sobre la eficacia de estas redes comparadas con modelos estadísticos tradicionales como Seasonal SARIMA, especialmente en la detección de anomalías vegetales en la región de estudio, con sus particularidades en cuanto a fertilidad, clima, presencia de sales, sodicidad, entre otras. Investigaciones como la de Malhotra et al. (2015) sugieren que, dependiendo de las condiciones específicas de la región de estudio, los métodos tradicionales podrían ser superiores, aunque es una afirmación que, debe ser sometida a prueba por las propiedades particulares de la región específica.

El trabajo de Atzberger, C. (2013) subraya la importancia de explorar modelos capaces de capturar relaciones no lineales, comunes en los datos de NDVI debido a las variaciones estacionales y eventos climáticos. Este estudio opta por el uso de LSTM como modelo de comparación, dado que el análisis de estas series temporales requiere una evaluación detallada para seleccionar el modelo más adecuado según las características específicas del área estudiada.

Debido a que la evolución de los cultivos exhibe potencialmente patrones estacionales y variaciones heterogéneas a lo largo del tiempo, la eficacia en la detección de anomalías depende en gran medida de la capacidad de los modelos para adaptarse a dichas características. En este contexto, se plantea una evaluación comparativa entre dos modelos estadísticos avanzados: SARIMA y LSTM. Ambos modelos son ampliamente reconocidos

Por su habilidad para capturar patrones estacionales y dinámicas complejas, los modelos LSTM se presentan como herramientas particularmente adecuadas para este tipo de análisis. La escasez de investigaciones previas sobre el comportamiento de ciertos modelos en la región del norte de Santa Fe y la falta general de conocimiento científico al respecto, subrayan la importancia de este estudio. Por lo tanto, el objetivo principal de esta investigación es evaluar la efectividad de los modelos LSTM en comparación con los modelos SARIMA para la detección temprana de anomalías en cultivos de soja en la región de estudio.

Cabe aclarar que Hj Mohd Rhymee et al. (2023), destaca que este análisis comparativo no sólo aclara las fortalezas y limitaciones de cada modelo en contextos específicos, sino que también orientará las futuras estrategias de monitoreo agrícola en la región, adaptando las tecnologías a las demandas particulares del sector.

Conforme a lo expuesto anteriormente, en el apartado 2 se presentan claramente los objetivos y sus respectivas hipótesis, que serán oportunamente resueltas en el capítulo 5.

2. Objetivos y Formulación de Hipótesis

2.1. Objetivo General.

Evaluar la efectividad de los modelos LSTM en comparación con los modelos SARIMA para la detección temprana de anomalías en cultivos de soja en la región de estudio.

2.2. Objetivos Específicos.

- Caracterización de la Dinámica Regional: Definir con precisión el área geográfica de estudio y describir la dinámica espacio-temporal de los cultivos de soja, enfocándose en el análisis de la evolución típica de los lotes.
- Evaluación de Modelos Predictivos: aplicar los modelos LSTM y SARIMA para evaluar su eficacia en capturar las dinámicas de las series temporales de NDVI, enfocándose en la comparación de su rendimiento en la detección temprana de anomalías de la región.
- Análisis de Precisión Predictiva y Aplicabilidad: Comparar la precisión predictiva de ambos modelos y examinar su aplicabilidad, considerando los posibles impactos prácticos en la gestión agrícola de la región.

A continuación, se definen las hipótesis de trabajo lo cual será puesta en prueba a lo largo del mismo, las cuales se encuentran fuertemente vinculadas a la detección de anomalías.

2.3. Hipótesis del trabajo.

- Las series temporales de NDVI, obtenidas de cultivos de desarrollo típico mediante sensores remotos, son fundamentales para identificar anomalías en el crecimiento de los cultivos agrícolas de forma automática.

- Los modelos SARIMA y LSTM resultan extremadamente útiles para identificar patrones en la evolución del ciclo de cultivos en la región de interés, y pueden emplearse eficazmente para la detección de anomalías en estos procesos.

- La utilidad de los modelos LSTM supera a los modelos SARIMA en la detección de anomalías en cultivos agrícolas, especialmente en la región estudiada.

3. Materiales.

3.1. Unidades de análisis y región de estudio.

Este estudio se centra en el campo perteneciente al campo denominado como San Santiago, el cual se encuentra localizado al noreste de la provincia de Santa Fe, específicamente se encuentra localizado en en las coordenadas 61°57'48.26" O, 29°56'54.20" S. Se emplearon datos cuantitativos obtenidos de imágenes satelitales y sensores remotos, resaltando la aplicación de herramientas modernas y de acceso libre para el procesamiento de dichas imágenes, como se menciona en la sección 3.4.3. Específicamente, se utilizó el sensor S2, descrito en la sección 3.4.2 del documento. Las imágenes proporcionadas por este sensor, disponibles desde 2015 y de dominio público, son esenciales para analizar los atributos de la vegetación en la región, y procesadas con Google Earth Engine (GEE). Ver apartado 3.6.3.

3.2. Fuente de datos.

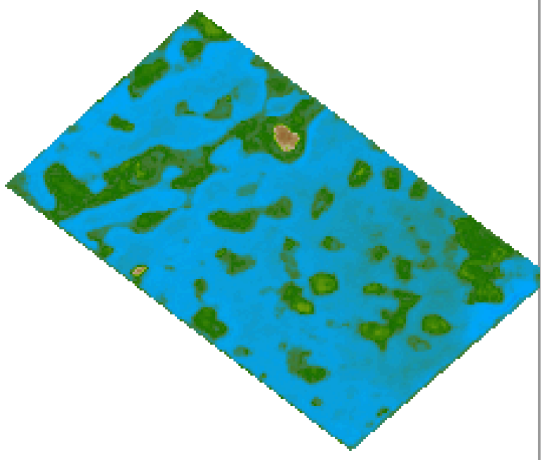
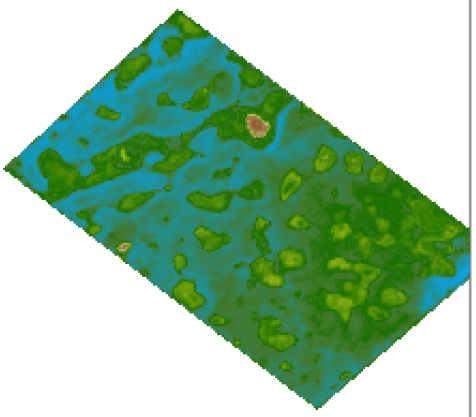
El estudio presenta una investigación cuantitativa que emplea técnicas avanzadas de análisis estadístico y aprendizaje profundo para investigar la dinámica de los cultivos de soja. Utiliza series temporales de NDVI derivadas de imágenes satelitales Sentinel-2 en la

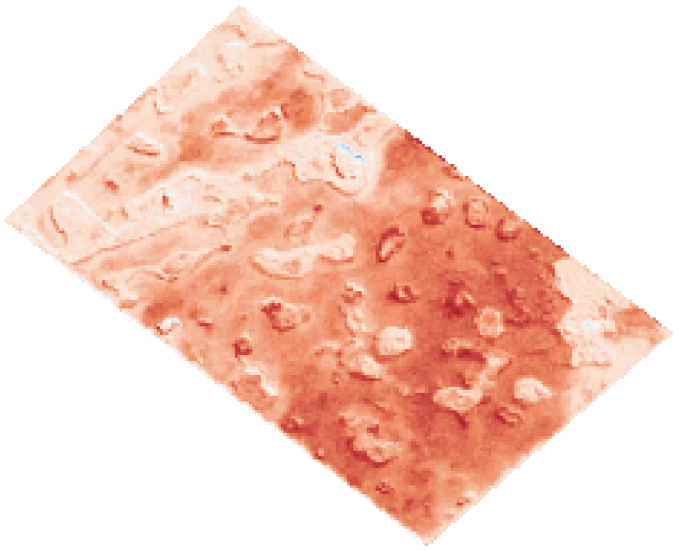
Región de Santa Fe, Argentina. La investigación se enriquece con datos complementarios provenientes de informes de campo elaborados por ingenieros agrónomos, lo que permite una verificación más rica de las observaciones satelitales.

A continuación, en la sección 3.3 del estudio, se realiza un detallado análisis y caracterización de la anomalía específica para el caso de estudio. Este segmento profundiza en la identificación y evaluación de patrones atípicos en las series temporales de NDVI, destacando su relevancia para comprender los impactos ambientales y de gestión sobre los cultivos de soja

3.3. Análisis y caracterización de la anomalía para el caso de estudio.

El informe detallado en la TABLA 1 ilustra una acumulación significativa de hectáreas que exhiben descensos en las tasas de cambio del NDVI, lo que indica una disminución en este índice. Este dato es un claro reflejo del impacto adverso de la helada reportada en la fecha indicada (correspondiente a 91 días después de la siembra). El deterioro asociado es también perceptible en las imágenes satelitales tomadas el 6 de abril de 2022, particularmente por la alteración de colores en la zona media-baja del terreno analizado.

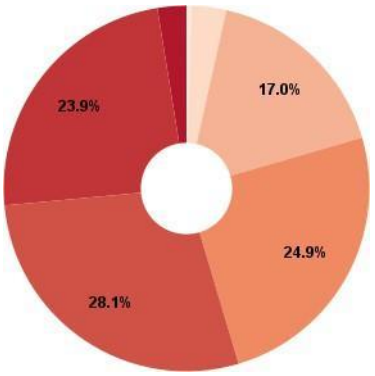
NDVI - 2022-04-01 Promedio: 0.79	NDVI - 2022-04-06 Promedio: 0.49	Simbología
		



La tasa de cambio expresa la variación diaria de cambio en porcentajes, variando entre -5.80% y + 5.80%.

Campaña: Gruesa 21-22 | **Establecimiento:** Cruz de Guerra | **Lote:** 09. Cultivo: **Soja de Segunda**

Distribución de Áreas según Tasa de Cambio



Tasa de cambio	Tasa de cambio diaria (Has)	Área Total (%)
de [3.980% a 4.870%)	0	0.0%
de [3.100% a 3.980%)	0	0.0%
de [2.210% a 3.100%)	0.021	0.0%
de [1.330% a 2.210%)	0.041	0.0%
de [0.440% a 1.330%)	0.155	0.1%
de [-0.440% a 0.440%)	0.681	0.3%
de [-1.330% a -0.440%)	6.48	3.1%
de [-2.210% a -1.330%)	35.753	17.0%

de [-3.100% a -2.210%)	52.345	24.9%
de [-3.980% a -3.100%)	59.145	28.1%
de [-4.870% a -3.980%)	50.323	23.9%
de [-5.750% a -4.870%)	5.376	2.6%

TABLA 1. Análisis de tasa de cambio para el Lote 09.

Un análisis más detallado de la variación del NDVI desde el 1 de abril de 2022, realizado a nivel de píxel utilizando imágenes del satélite Sentinel-2 (S2), muestra una disminución promedio del NDVI de 0.81 a 0.49. Este cambio confirma el daño parcial sufrido por el lote, como se desprende de la interpretación de la paleta de colores utilizada. Además, se identifica una tendencia negativa en las tasas de cambio durante este periodo. Conforme al cronograma de fechas de siembra y cosecha, detallado en la FIGURA 1, esta caída es inesperada para el intervalo observado.

Se implementaron dos análisis utilizando el método K-Medias, ambos con un valor de K=3. El primer análisis corresponde a la fecha del 1 de abril de 2022, y el segundo al 6 de abril de 2022. Este enfoque permite una visualización diferenciada y detallada del impacto en las áreas analizadas.

Referencias:

- A: NDVI Alto.
- M: NDVI Medio.

- B: NDVI Bajo.

Esta estructura proporciona una presentación clara de las fechas de análisis y las categorías de NDVI utilizadas, facilitando la comprensión de los resultados obtenidos.

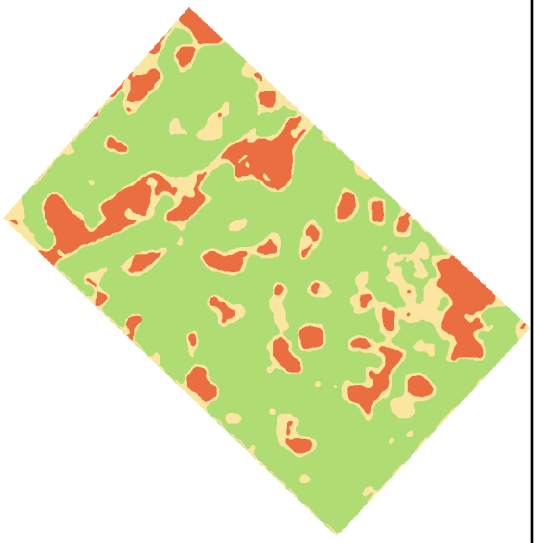
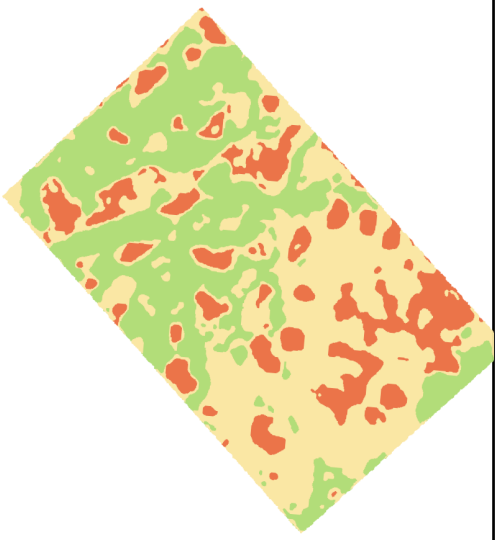
K medias K=3 fecha 2022/04/01	K medias K = 3 fecha 2022/04/06	Referencia
		
Referencias: **A**: NDVI Alto. **M**: NDVI Medio. **B**: NDVI Bajo.		

TABLA 2. Análisis de clustering para el Lote 09 entre las fechas 2022/04/01 y 2022/04/06

El método de K-Medias, junto con otros enfoques detallados en la Tabla 4, ha identificado un declive en el NDVI en varios píxeles del lote, un fenómeno que contradice las expectativas de crecimiento del cultivo y que claramente se relaciona con el desastre reportado anteriormente en esas fechas.

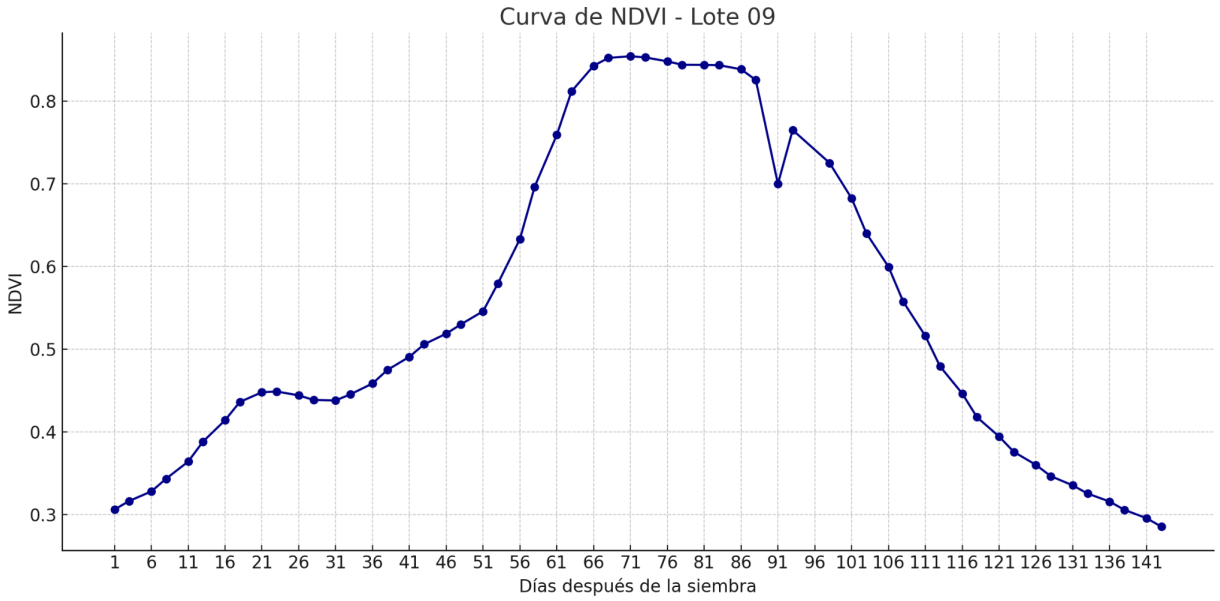


FIGURA 1. Serie temporal de NDVI para el lote 09.

Habiendo descrito e identificado la anomalía, y definido claramente los materiales y unidades de análisis empleados, el siguiente paso consiste en realizar una revisión conceptual. Esta revisión es fundamental para asegurar una comprensión adecuada y la correcta aplicación de los modelos analíticos en el estudio. Como se observa en la Figura 1, la anomalía detectada es consistente con lo descrito en la sección 3.2.2 de este trabajo, que trata específicamente de una anomalía puntual. Observe que la anomalía (dia 91 después de la siembra), ha sido reducida de 0.7 a valores más altos, por la suavización misma.

3. Revision conceptual.

3.1. Definiciones útiles en series temporales.

3.1.1. Series temporales.

Una serie de tiempo consiste en un conjunto de observaciones ordenadas temporalmente, sobre un fenómeno dado (variable objetivo). Por lo general, las mediciones están igualmente espaciadas, por ejemplo, anualmente, trimestralmente, mensualmente, semanalmente o diariamente. La propiedad más importante de una serie temporal es que las observaciones ordenadas son dependientes del tiempo, y la naturaleza de esta dependencia es de interés en sí misma. Formalmente, una serie temporal se define como un conjunto de variables aleatorias indexadas en el tiempo, $\{Y_1, \ldots, Y_t\}$. el subíndice t indica el tiempo.

La aplicación de métodos de series temporales permite descomponer la serie en sus componentes fundamentales: tendencia, efectos estacionales e irregularidades. Según Enders et al. (2014), este enfoque facilita un análisis más profundo y una mejor comprensión de la estructura subyacente de la serie temporal.

3.1.2. Tendencia.

En el análisis de series temporales, la tendencia describe un patrón persistente que refleja crecimiento o decrecimiento en los datos a lo largo del tiempo. Este patrón señala la dirección general que siguen los datos y es esencial para entender comportamientos en ámbitos como la economía, la meteorología y lo social.

Las tendencias pueden variar en forma, no limitándose a patrones lineales; pueden ser exponenciales, logísticas, polinomiales o incluso más complejas. Cada tipo de tendencia aporta conocimientos distintos sobre el comportamiento del fenómeno estudiado. Por ejemplo, una tendencia exponencial indica un crecimiento acelerado, mientras que una lineal muestra un aumento constante.

Identificar la tendencia es clave para el análisis predictivo y la toma de decisiones estratégicas. Comprender su dirección y forma permite a los analistas hacer proyecciones, ajustar estrategias y prever cambios en los patrones observados. Además, reconocer tendencias facilita la descomposición de la serie temporal en componentes más manejables, mejorando el análisis y la interpretación de los datos.

3.1.3. Estacionalidad.

Según Hyndman et. al (2021), la estacionalidad es un componente esencial en el análisis y modelado de series temporales. Este término describe las fluctuaciones periódicas y predecibles que se presentan en ciclos fijos, influenciadas por factores estacionales como la época del año, el día de la semana o la hora del día. La regularidad y previsibilidad de la estacionalidad permiten a investigadores y analistas identificar y modelar estas variaciones, lo cual mejora la precisión de las predicciones y el entendimiento de la serie temporal.

3.1.4. Estacionariedad.

En matemáticas, un proceso estacionario o estrictamente estacionario es un tipo de proceso estocástico cuya distribución de probabilidad permanece constante a lo largo del tiempo o en una posición específica. Esto implica que parámetros estadísticos como la media y la varianza, si existen, se mantienen invariables con el tiempo o la posición.

Un proceso puede manifestar no estacionariedad en aspectos como la media, la varianza, las autocorrelaciones o en la distribución general de las variables. Por ejemplo, una serie no estacionaria en la media si su nivel no es constante a lo largo del tiempo, mostrando posibles tendencias ascendentes o descendentes. Si la variabilidad o las autocorrelaciones cambian con el tiempo, la serie es no estacionaria en la varianza o en las autocorrelaciones. Además, si la distribución de las variables cambia en cada momento, la serie es no estacionaria en la distribución.

Los procesos no estacionarios relevantes incluyen los procesos integrados, que se caracterizan por convertirse en estacionarios una vez diferenciados.

Para los modelos de predicción, es crucial trabajar con series estacionarias para obtener proyecciones confiables. Por esta razón, la prueba de estacionariedad es una práctica común en el análisis de series temporales.

En síntesis, estrictamente un proceso es estacionario cuando:

1. Las distribuciones marginales de todas las variables son idénticas (Distribuciones idénticas para cada retardo).
2. Las distribuciones finito-dimensionales de cualquier conjunto de variables solo

Matemáticamente: $F(y_i, y_j, \ldots, y_k) = F(y_{i+h}, y_{j+h}, \ldots, y_{k+h})$. Esta propiedad suele denominarse estacionariedad fuerte; sin embargo, es difícil de demostrar. Por esta razón, en este trabajo nos centramos en la estacionalidad débil, manteniendo, no obstante, el interés en la estacionalidad fuerte. A diferencia de la estacionariedad fuerte, la estacionalidad débil implica constancia en la media, la varianza y la estructura de covarianzas a lo largo del tiempo.

Matemáticamente, un proceso es estacionario en el sentido débil si, para todo instante t, se cumple que:

1. $E(y_t) = u$, con varianza constante.
2. $\Upsilon(t, t-k) = E[(y_t - \mu)(y_{t-k} - \mu)] = \Upsilon_k \quad k = -1, -2, 0, 1, 2.$

Adicionalmente, en este contexto, es de interés mencionar el proceso de ruido blanco, el cual es un proceso estacionario que desempeña un papel crucial en la predicción de series temporales. Este proceso se define como un conjunto de valores aleatorios en el que las observaciones en diferentes tiempos son estadísticamente independientes, es decir, no presentan correlación entre sí. En el contexto de análisis de series temporales, el ruido blanco es especialmente valioso para examinar los residuos de los modelos. Si los residuos siguen un patrón de ruido blanco, generalmente se considera que el modelo tiene un buen ajuste. Cabe aclarar que esto no es determinante. Hay diferentes pruebas estadísticas que pueden realizarse para determinar, con cierta probabilidad, si una serie se corresponde a un ruido blanco. por ejemplo, la prueba de Ljung-Box.
Estos procesos, responden a las siguientes propiedades:

- $E(y) = 0$ proceso para el cual la media es igual a 0.
- $\text{Var}(y) = \text{constante}$.
- $Cov(y, y_{t+h}) = 0$ proceso para el cual todas sus variables son independientes.

Es probable que un proceso aleatorio no se corresponda con un ruido blanco si se cumple al menos alguna de las siguientes condiciones:

- La media no es igual a 0.
- La media cambia a lo largo del tiempo.
- La varianza cambia a lo largo del tiempo.
- Los valores están correlacionados con periodos anteriores.

FIGURA 2. Ejemplo gráfico de un proceso de ruido blanco.

3.1.5. Correlación Serial y Función de Autocorrelación Simple (FAC).

Las autocorrelaciones contienen la misma información que las autocovarianzas, con la ventaja de no depender de las unidades de medida.

La correlación serial es una medida que indica si los valores que toma una variable en el tiempo no son independientes entre sí, sino que un valor determinado depende de los anteriores de la misma serie o de otras. La estructura de dependencia lineal entre las variables aleatorias se representa por las funciones de covarianza y correlación. Se puede decir que la correlación es una medida de la relación lineal entre variables, específicamente, mide la relación lineal entre los retrasos en series temporales. Hay varios coeficientes de correlación, correspondientes a cada panel en el gráfico de autocorrelación.

Por ejemplo, r_1 mide la relación entre y_t y y_{t-1}, r_2 la relación entre y_t y y_{t-2}, y así sucesivamente.

El valor r_k puede ser escrito como:

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=k+1}^{T} (y_t - \bar{y})^2}.$$

La función de autocorrelación mide la correlación entre dos variables separadas por un cierto número de periodos. La dependencia temporal es un fenómeno muy habitual en las series de tiempo.

La función de autocorrelación mide la correlación entre dos variables separadas por un cierto número de periodos. La dependencia temporal es un fenómeno muy habitual en las series de tiempo.

3.1.6. Función de autocorrelación parcial (FAPC).

En el análisis de series de tiempo, la FAPC desempeña un papel crucial en el análisis de datos orientado a identificar la medida de desfase en modelos autorregresivos. La incorporación de esta función es una parte fundamental de la metodología de Box-Jenkins para la modelación de series temporales, donde mediante el trazado de las funciones de autocorrelación parciales se podría determinar los rezagos apropiados p en un modelo $AR(p)$ o en uno $ARIMA\ (p, d, q)$.

Determinar el orden de un proceso autorregresivo utilizando su función de autocorrelación simple puede ser desafiante. Para enfrentar este reto, se utiliza la función de autocorrelación parcial (FACP), que mide la correlación entre una observación en el

tiempo t y otra en el tiempo $t - k$ controlando por las observaciones intermedias. Esta función resulta especialmente útil cuando se analizan datos con tendencias, dado que en estos contextos la autocorrelación para rezagos cortos tiende a ser alta y positiva.

3.1.7. Modelo autorregresivo integrado de media móvil (ARIMA).

El modelo consiste en la combinación de un término autorregresivo (AR) y un término de promedio móvil (MA), complementado con un elemento de diferenciación representado por la letra "I". Este componente se basa en un estudio realizado por Yaglom (1955).
En general estos modelos se referencian como *ARIMA*(p, d, q). Donde el parámetro p se refiere al orden del modelo autorregresivo; d, al término de diferenciación, y q, al término de media móvil con q términos de error.

De forma precisa, si d es un entero no negativo, entonces $\{X_t\}$ es un proceso $ARIMA(p,d,q)$ si $Y_t = (1-B)^b X_t$ es un proceso $ARMA(p,q)$ *de acuerdo con* de acuerdo con Brockwell y Davis (2002) a continuación, se introduce el modelo $SARIMA$, el cual resulta objeto del presente trabajo.

3.1.8. Modelo SARIMA.

El modelo $SARIMA$ Ha sido diseñado para capturar patrones estacionales en los datos de series temporales. La notación $SARIMA(p,d,q)x(P,D,Q)_s$ describe tanto los componentes no estacionales como los componentes estacionales del modelo, los mismos son descritos a continuación:

- p orden AR no estacional.
- d orden de diferencia no estacional.
- q orden MA no estacional.
- P orden AR estacional.
- D orden de diferencia estacional.
- Q orden MA estacional.
- s ventana de tiempo del patrón estacional.

Sin operaciones de diferenciación, el modelo podría expresarse matemáticamente como:

$\Phi(B^s)\phi(B)(x_t - \mu) = \Theta(B^s)\theta(B)w_t$, donde:

Los componentes no estacionales son:

- AR no estacional $\phi(B^s) = 1 - \phi_1 B - \ldots - \phi_p B^p$.
- MA no estacional $\theta(B^s) = 1 + \theta_1 B + \ldots + \theta_q B^q$.

Los componentes estacionales son:

- AR estacional $\Phi(B^s) = 1 - \Phi_1 B^s \ldots \Phi_P B^{Ps}$.
- MA estacional $\Theta(B^s) = 1 + \Theta_1 B^s + \ldots + \Theta_Q B^{Qs}$.

3.1.9. Predicciones con modelo SARIMA.

Según Brockwell et al. (2002), las predicciones generadas mediante modelos SARIMA son completamente análogas a las producidas por los modelos ARIMA previamente revisados. No obstante, las predicciones carecen de valor práctico si no se acompañan de una medida de su precisión. Por esta razón, en el presente trabajo se analiza la distribución de los errores de predicción, con el fin de evaluar la capacidad predictiva de los modelos utilizados.

3.2. Anomalías en series de tiempo.

En este apartado, comenzamos a definir qué se entiende por anomalía en series de tiempo, y realizamos una categorización de las mismas realizando comentarios sobre la aplicación de la misma al caso de estudio.

3.2.1. Anomalía.

Uno de los conceptos clave abordados en este trabajo es el de anomalía. De acuerdo con Chandola, et. al (2009), una anomalía es un patrón en los datos que no se ajusta al comportamiento esperado; también se la denomina *outlier*, excepción, peculiaridad o sorpresa, entre otros términos. La detección y análisis de anomalías puede abordarse de diferentes maneras, según el tipo de anomalía presente. Por ello, a continuación se revisan las principales categorías de anomalías en series temporales. Se presenta una breve taxonomía que permite clasificar los distintos tipos de comportamientos anómalos que pueden observarse al analizar este tipo de datos.

3.2.2. Tipos de anomalías en series temporales.

Las técnicas de detección de anomalías varían en función del tipo de datos analizados, la naturaleza de las anomalías y los métodos empleados para su identificación. Según Blázquez-García, et. al (2021), es posible establecer una taxonomía específica para clasificar estas técnicas, la cual se presenta en la FIGURA 3.

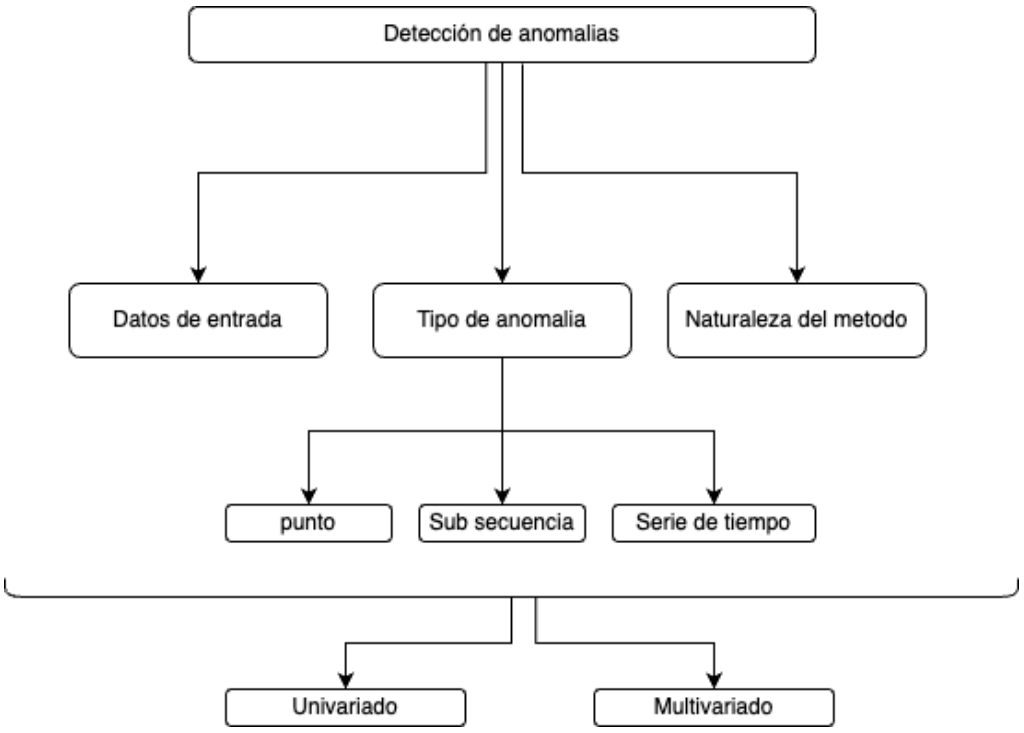


FIGURA 3. Taxonomía de las técnicas de detección de anomalías en series temporales. Blázquez et al. (2021).

Las anomalías en series temporales pueden clasificarse según la taxonomía de la FIGURA 3, es importante destacar que las mismas varían en función de los datos de entrada, el tipo de anomalía. En este trabajo, se emplea una serie de tiempo univariada, que se caracteriza por analizar una única variable a lo largo del tiempo.

Según Blázquez et al. (2021), se identifican tres tipos principales de anomalías: puntuales, subsecuencias anómalas y series temporales anómalas. Una anomalía puntual se detecta cuando una única medición difiere significativamente del comportamiento típico observado en el resto de la serie. Las subsecuencias anómalas se caracterizan por grupos de puntos consecutivos que presentan comportamientos atípicos. Por otro lado, las series

temporales anómalas, que generalmente requieren un análisis multivariado, se definen por patrones inusuales que se mantienen a lo largo de toda la serie.

En este estudio, se examina una serie temporal univariada para identificar tanto subsecuencias de puntos anómalos como series temporales anómalas, lo que permite profundizar en la comprensión y detección de estos fenómenos atípicos.

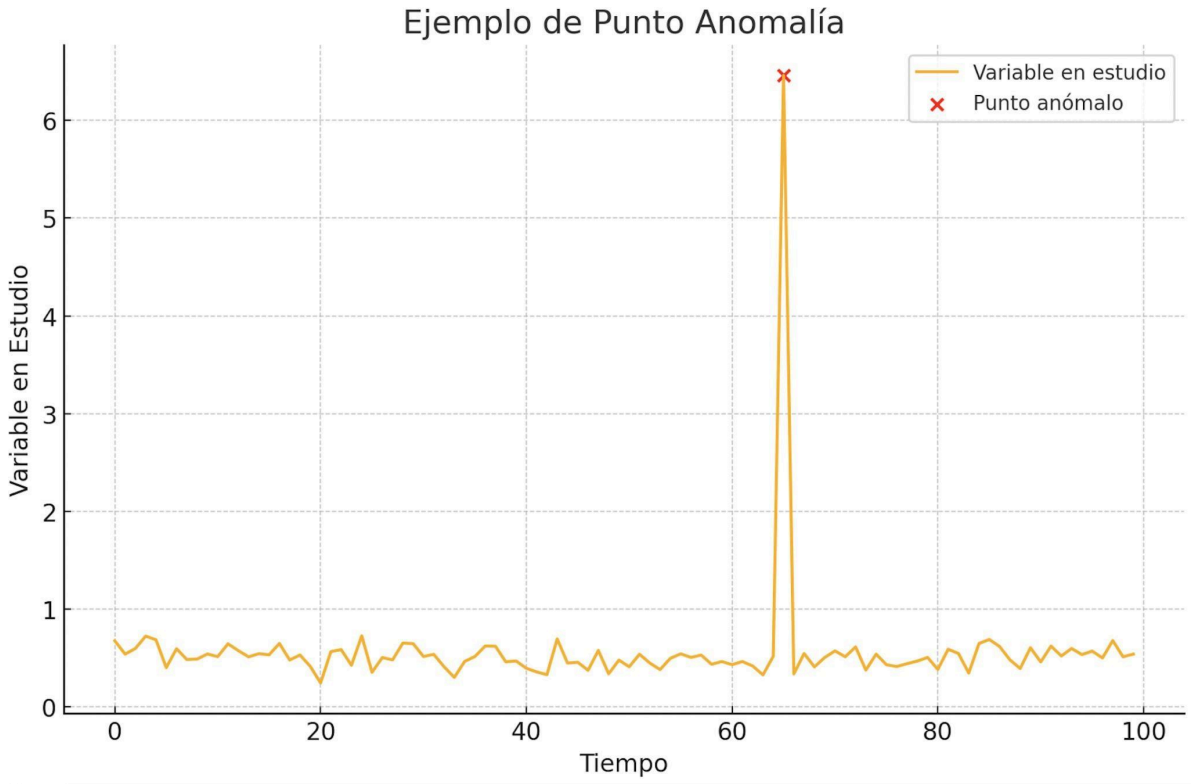


FIGURA 4. Ejemplo de punto anómalo de propia autoría.

Según la FIGURA 4, se identifica un instante de tiempo anómalo como aquel en el que se observan valores inusuales en comparación con el resto de la serie o en relación a observaciones de momentos cercanos, lo cual puede ser una anomalía global o local, respectivamente. Un instante de tiempo anómalo puede ser univariante o multivariante, dependiendo de si afecta a una o varias variables de la serie.

Este estudio se centra en el análisis univariado, empleando exclusivamente la variable NDVI. Este enfoque se dedica a examinar la evolución de una sola variable a lo largo del tiempo. En el contexto de la detección de anomalías, se adoptan modelos especializados para el análisis de residuos y la identificación de valores atípicos. el Autor Akouemo et al. (2017) implementan modelos SARIMA con variables exógenas, regresión lineal y redes LSTM para fines similares. Aunque estos modelos también se emplean en tareas de

predicción, en este caso particular, se centran en la detección de anomalías mediante la metodología apartado 4, básicamente se utilizan utilizando tanto datos históricos como proyecciones futuras con la propuesta metodológica en cuestión. Es importante mencionar que este trabajo excluye las variables exógenas debido a los desafíos en la recolección de datos adicionales y la complejidad inherente al NDVI, el cual como se menciona en la introducción, es una variable que resulta útil para el estudiante.

Anomalías en el campo de la teledetección y agricultura.

Las anomalías en imágenes de satélite se definen como cambios significativos en la cobertura terrestre o disturbios causados por factores naturales, tales como incendios, inundaciones, sequías severas y tormentas, o por factores biogénicos como enfermedades de plantas e infestaciones de insectos. Este estudio se enfoca en la identificación de estas perturbaciones mediante índices derivados del procesamiento de imágenes satelitales, clasificando cualquier observación o patrón que se aparte del comportamiento esperado como una anomalía. Blázquez et al. (2020) y Li y Jung (2023) señalan que existen diversos tipos de anomalías que pueden detectarse en series temporales.

Es esencial entender la terminología y las metodologías de los dos modelos seleccionados para la comparación. La comprensión de estos fenómenos es crucial, especialmente en la agricultura, donde variables como el NDVI y la fecha de siembra juegan un papel fundamental. Las variaciones regionales en condiciones ambientales y climáticas, como la temperatura, precipitación y radiación solar, impactan directamente en el crecimiento y los ciclos fenológicos de los cultivos.

Una referencia significativa es el estudio de Moreno et al. (2018), que examina cómo las variaciones regionales del NDVI afectan la modelización de la productividad agrícola bajo distintas condiciones climáticas. A partir de estos hallazgos, se ajustan los modelos para utilizarlos como puntos de referencia y evaluar su efectividad en la detección de anomalías en la región estudiada. Esta área presenta características climáticas únicas que influyen en la evolución temporal de los cultivos y, por ende, en los resultados obtenidos por los modelos.

3.3. Definiciones útiles para redes neuronales.

Esta sección presenta el concepto de redes neuronales e introduce los modelos de redes neuronales adecuadas para el modelo de series temporales.

3.3.1. Aprendizaje supervisado.

Raschka et al. (2019) clasifican el aprendizaje automático en tres categorías principales: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. En el contexto de este estudio, se utiliza exclusivamente el aprendizaje supervisado para ajustar algoritmos en redes neuronales. Este enfoque se caracteriza por emplear datos previamente etiquetados en tareas de clasificación o conjuntos de datos numéricos específicos para regresión, los cuales se utilizan en el entrenamiento del modelo.

3.3.2. Pronósticos con aprendizaje supervisado.

El aprendizaje supervisado tiene como objetivo principal desarrollar un modelo que pueda clasificar o realizar predicciones sobre datos futuros o desconocidos, partiendo de datos de entrenamiento que ya han sido etiquetados o observados. En este contexto, el término 'supervisado' se refiere al empleo de un conjunto de muestras cuyas respuestas o etiquetas deseadas ya son conocidas, lo cual facilita significativamente la tarea de clasificación o regresión.

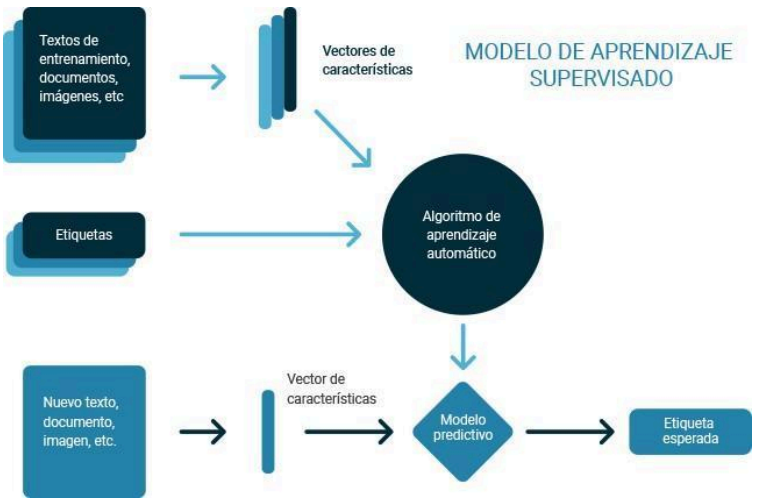


FIGURA 5. Aprendizaje Supervisado

Como se resume en la FIGURA 5, el aprendizaje supervisado ofrece un marco robusto para modelar tanto problemas de clasificación como de regresión. En la regresión, a diferencia de la clasificación que predice categorías, el objetivo es predecir un valor continuo. Esto se ilustra cuando se introducen datos de entrada junto con sus correspondientes valores continuos o etiquetas de salida conocidos. Estos elementos son procesados por el algoritmo, que mediante un proceso iterativo de ajuste busca minimizar la diferencia entre los valores predichos y los reales, convergiendo hacia una solución óptima. Una vez ajustado, el modelo puede aplicarse a nuevos datos de entrada, proporcionando predicciones precisas para valores continuos, asumiendo que el modelo está bien optimizado.

Aunque la predicción de series temporales se distingue en ciertos aspectos de los problemas de regresión típicos, comparte similitudes metodológicas. En particular, técnicas de regresión se aplican dentro de métodos de pronóstico de series temporales, como en los modelos autorregresivos (AR) y los de media móvil (MA), que se basan en relaciones continuas para predecir valores futuros. Los modelos de series temporales también pueden ser enriquecidos con variables exógenas para aumentar la precisión de las predicciones, mostrando la versatilidad y la conexión entre el análisis de series temporales y la regresión.

3.3.3. Redes neuronales.

La investigación en redes neuronales artificiales, a menudo simplemente llamadas "redes neuronales", se ha inspirado en el entendimiento de que el cerebro humano opera de manera radicalmente distinta a la de las computadoras digitales tradicionales. A diferencia de estos sistemas convencionales, el cerebro se caracteriza por ser un procesador altamente complejo, no lineal y operando en paralelo. Este sistema de procesamiento de información destaca por su habilidad para organizar sus neuronas en la ejecución de tareas específicas, tales como el reconocimiento de patrones, la percepción sensorial y el control motor.

Una red neuronal artificial se conceptualiza como un procesador masivamente paralelo y distribuido, integrado por unidades de procesamiento simples, diseñado no solo para almacenar conocimientos empíricos, sino también para facilitar su aplicación práctica.

Las redes neuronales artificiales se asemejan al cerebro humano en dos aspectos fundamentales: primero, estas redes adquieren conocimientos de su entorno mediante un proceso de aprendizaje; segundo, la información aprendida se almacena a través de la variación en la fuerza de las conexiones entre interneuronas, denominadas pesos

sinápticos.

Para ilustrar, una neurona artificial puede ser representada como se muestra en la siguiente ilustración.

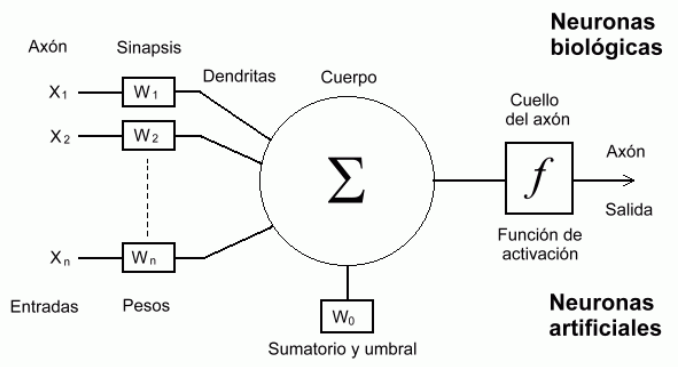


FIGURA 6. Perceptrón simple.

Este modelo es conocido como perceptrón simple y es el modelo base para construir redes neuronales multicapas. La neurona de la FIGURA 6, replica el comportamiento de la neurona biológica.

La imagen muestra una unidad neuronal, con un vector de entrada $\{x_1, \ldots, x_n\}$, los pesos de conexión $\{W_1, \ldots, W_n\}$ los cuales son ajustados conforme avanza el proceso de entrenamiento y finalmente un parámetro de bias, el cual permite que las soluciones (rectas, planos, hiperplanos de separación) no sean forzadas al origen de coordenadas.

Las entradas se multiplican por sus respectivos pesos y se combinan mediante una suma. A partir de allí, se genera una salida que pasa por una función de activación $f(x) = \frac{1}{1+e^{-x}}$.

Las redes neuronales se forman a través de la combinación de múltiples unidades neuronales. Uno de los tipos más comunes es la red neuronal de perceptrón multicapa. A continuación, presentamos un ejemplo de este tipo de red y analizamos los algoritmos asociados, que son fundamentales para entender las Redes Neuronales Recurrentes (RNN), las cuales son de especial interés en este estudio.

3.3.4. Redes neuronales de perceptrón multicapas.

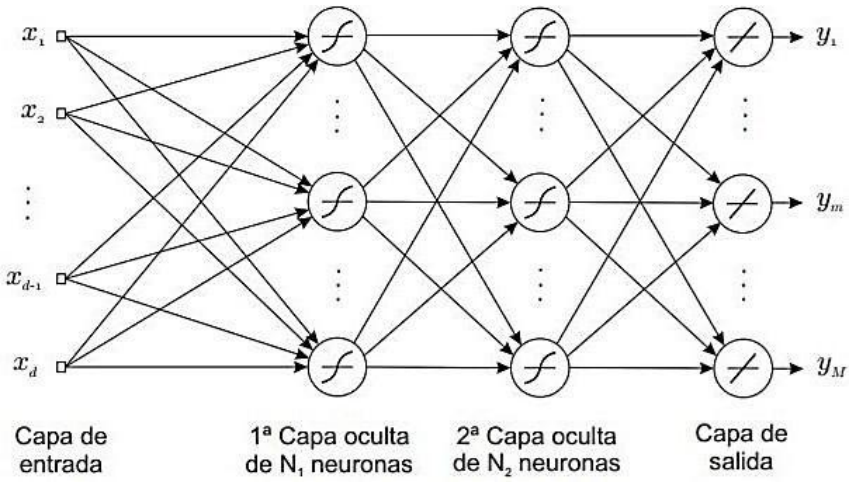


FIGURA 7. Red completamente conectada, también llamadas redes multicapas.

Como muestra la FIGURA 7, un conjunto de perceptrones que se conectan pueden modelarse con una red neuronal multicapa, o perceptrón multicapa. Los perceptrones multicapa se han aplicado con éxito para resolver algunos problemas difíciles, tanto de clasificación como de regresión, tienen la característica de resolver problemas no lineales, los cuales son una limitante para el perceptrón simple, la sección 3.3.3.

Se caracteriza por tener una capa de entrada, n capas ocultas, y una capa de salida, donde finalmente es computada la predicción. Las conexiones entre neuronas se caracterizan por tener pesos, perteneciente al conjunto numérico de los reales.

Su entrenamiento (ajuste de los parámetros o pesos) se realiza de manera supervisada, y utilizando generalmente el algoritmo de propagación hacia atrás o descenso del gradiente, este algoritmo se puede consultar en la sección 3.3.5.

Básicamente, el aprendizaje de la propagación hacia atrás del error consiste en dos pases a través de las diferentes capas de la red: *un pase hacia adelante* y *un pase hacia atrás*. En el pase hacia adelante, se aplica un patrón de actividad (vector de entrada) a los nodos sensoriales de la red, y sus efectos se propagan a través de la red capa por capa. Finalmente, se produce un conjunto de resultados como la respuesta real de la red. Durante el paso de avance, los pesos sinápticos de las redes son todos fijos. Durante el pase hacia atrás, por otro lado, los pesos sinápticos se ajustan todos de acuerdo con una regla de corrección de errores, por ejemplo, mediante el conocido y documentado algoritmo de Backpropagation.

3.3.5. Algoritmo de Propagación hacia atrás (Backpropagation).

Uno de los algoritmos más utilizados para entrenar redes neuronales multicapa es el algoritmo de propagación hacia atrás. Este término, "back-propagation", se popularizó tras su aparición en 1985 y se hizo conocido principalmente a través de la publicación *Parallel Distributed Processing* de Rumelhart et al. (1986). El desarrollo de este algoritmo marca un hito en el campo de las redes neuronales, ya que ofrece un método computacionalmente eficiente para el entrenamiento de perceptrones multicapa. Aunque no se puede afirmar que la retropropagación sea una solución óptima para todos los problemas resolubles, sí ha ayudado a disipar el pesimismo sobre el aprendizaje en máquinas multicapa, que pudo haberse inferido de la obra de Minsky y Papert (1969). El algoritmo se divide esencialmente en dos fases: una fase de computación hacia adelante basada en el dato de entrada, y una fase de propagación hacia atrás de la salida computada previamente.

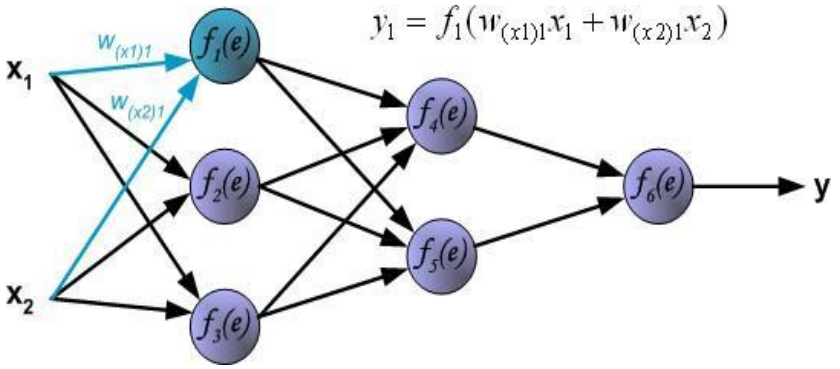


FIGURA 8. Ejemplo de la computación hacia adelante.

La computación hacia adelante permite, partiendo de un vector de entrada **x**, obtener el vector de respuesta **y**, ver Figura 8 de ejemplo (cálculo de y_1). La dimensión de **y** se determina por la cantidad de neuronas en la capa de salida. Por ejemplo, en problemas de clasificación que involucran tres categorías posibles de resultados $\mathbf{y} \in \Re^3$. En cuanto a la selección de la función f_1, típicamente se opta por una función que sea continua y diferenciable, ya que esto facilita el cálculo del gradiente de la función compuesta respecto a los pesos de las entradas de cada neurona (utilizando la regla de la cadena para el cálculo de las derivadas parciales). Es importante destacar que el algoritmo tiene como objetivo

principal minimizar el error mediante el método de descenso del gradiente conforme se avanza en lo llamado etapas o iteraciones de entrenamiento, utilizando para ello el algoritmo de retropropagación. Un ejemplo de función utilizada en este contexto es la función sigmoide.

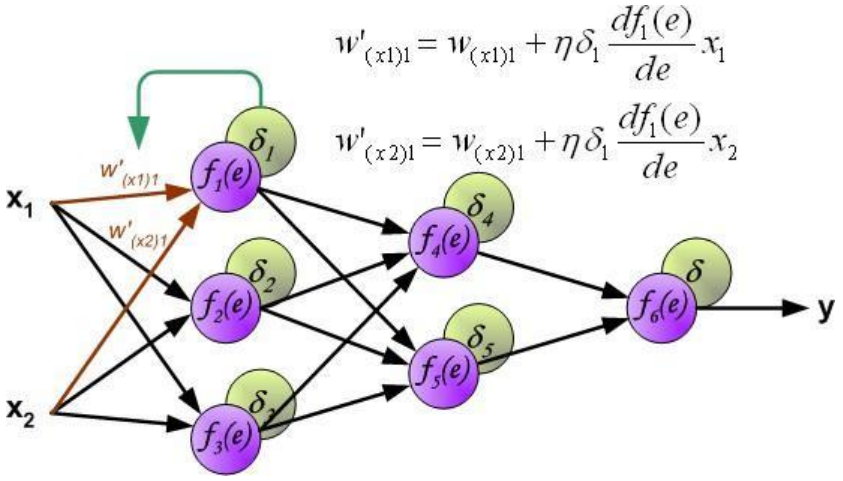


FIGURA 9. Ejemplo esquemático de propagación del error.

El ejemplo previo en la FIGURA 9 ilustra el ajuste de los pesos para la conexión específica con la primera neurona de la capa de entrada. Este proceso implica realizar cálculos específicos y ajustar los pesos de manera iterativa, con el objetivo de minimizar el error cometido por la misma. Buscamos así encontrar la configuración óptima de pesos **W** que permita una aproximación más precisa de la función deseada, ya sea para tareas de clasificación o de regresión. Estos pesos **W** son aquellos que minimizan la función de pérdida o error.

salida. Si esta salida refleja un error, el algoritmo responde actualizando los pesos de manera adecuada para corregir dicho error. En su trabajo, Joaquín Bermejo. (2022). Simplificó un algoritmo de descenso del gradiente, Este algoritmo explica cómo se ajustan las redes neuronales multicapa en función de los datos proporcionados. De manera simplificada, los pesos almacenados en el vector W se ajustan para minimizar la función de costo C. El algoritmo busca el mínimo en la dirección opuesta al gradiente, y el ajuste depende de la tasa de aprendizaje, la cual es seleccionada experimentalmente por el experto. Es relevante destacar que tanto la situación inicial de los pesos como los términos de bias se inicializan con valores aleatorios, que usualmente oscilan en el rango de [-1, 1]. A partir de lo expuesto, se definen estructuras más sofisticadas como las redes neuronales

recurrentes. A continuación, se presenta otro tipo de redes neuronales que, aunque comparten similitudes con las anteriormente descritas, son las que se utilizan en este trabajo.

Algoritmo 1: Descenso de gradiente

Datos: conjunto de observaciones $\mathbf{X}|\mathbf{Y}$, función de costo C, tasa de aprendizaje η

$w_{jk}^{l} \leftarrow \mathcal{U}(-1,1) \quad \forall\, j = \overline{1, n_l} \quad \forall\, k = \overline{1, n_{l-1}} \quad \forall\, l = \overline{2, L}$

$b_{j}^{l} \leftarrow \mathcal{U}(-1,1) \quad \forall\, j = \overline{1, n_l} \quad \forall\, l = \overline{2, L}$

para $e = \overline{1, E}$ **hacer**

  para $l = \overline{2, L}$ **hacer**

    $\mathbf{b}^{l} \leftarrow \mathbf{b}^{l} - \eta \dfrac{\partial C}{\partial \mathbf{b}^{l}}$

    $\mathbf{W}^{l} \leftarrow \mathbf{W}^{l} - \eta \dfrac{\partial C}{\partial \mathbf{W}^{l}}$

Resultado: $\{\mathbf{W}^{l}\}_{l=2}^{L}, \{\mathbf{b}^{l}\}_{l=2}^{L}$

FIGURA 10 . Algoritmo de descenso de gradiente Joaquín Bermejo. (2022).

3.3.6. Redes neuronales recurrentes (RNN).

Las Redes neuronales recurrentes, son una familia de redes neuronales para el procesamiento de datos secuenciales. Así como las redes neuronales convolucionales, que están especializadas en el procesamiento de grillas de valores X, tal como imágenes, una red neuronal recurrente se encuentra especializada para el procesamiento de secuencia de valores $\{x^{(1)}, .., x^{(t)}\}$. La mayoría de las redes neuronales recurrentes pueden también procesar secuencias de longitud variable. Rumelhart et al., (1986).

Las RNN pueden ser consideradas como sistemas dinámicos; se caracterizan por tener un estado interno en cada paso del tiempo en la clasificación. Esto es debido a una conexión circular entre capas de neuronas altas y bajas, y opcionalmente, conexiones hacia ellas mismas (retroalimentación), estas conexiones las dota de memoria y las hace adecuadas para la modelización de series temporales. A través del bucle continuo de información, las propias salidas de la red se convierten en entradas de instantes posteriores.

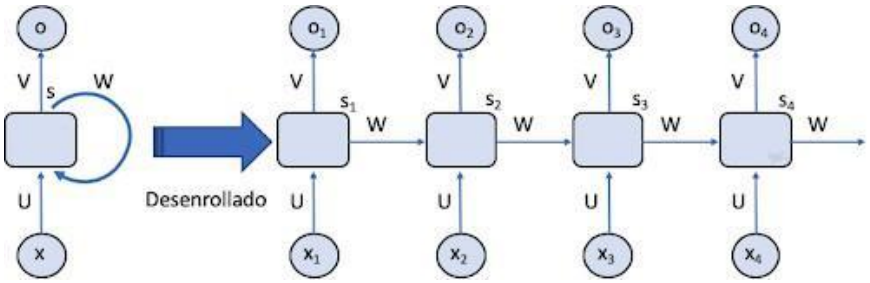


FIGURA 11. Redes neuronales recurrentes.

Las RNN tienen todas las características de las redes neuronales sencillas, con la adición de entradas que reflejan el estado de la iteración anterior. En cada iteración t, la salida contiene información de las $t-1$ iteraciones anteriores, a través de la entrada a la red del estado anterior, que a su vez tiene información del estado anterior a este, y así sucesivamente.

La información de salida en la iteración t alimentará a la red en la iteración $t+1$. En cada iteración la variable de estado es tratada como una variable de entrada más. Gutiérrez, E. M. (2019).

La imagen muestra una RNN desplegada en el tiempo. Una RNN es una red neuronal que tiene conexiones recurrentes, lo que significa que la salida de una capa se retroalimenta a la entrada de la misma capa. Esto permite que la red aprenda secuencias de datos.

En la Figura 11, la red neuronal recurrente tiene una capa de entrada, una capa oculta y una capa de salida. La capa de entrada recibe la entrada de datos, la capa oculta procesa los datos y la capa de salida produce la salida de la red.

La red neuronal recurrente se despliega en el tiempo, lo que significa que la misma red se utiliza para procesar cada elemento de la secuencia de datos. En la imagen, la red se despliega cuatro veces, una vez para cada elemento de la secuencia de datos.

La salida de la red neuronal recurrente es una secuencia de valores. En la imagen, la salida de la red es una secuencia de cuatro valores, uno para cada elemento de la secuencia de datos.

Estas RNN tienen sentido del tiempo y de la memoria a corto plazo, ya que la memoria a largo plazo en realidad es prácticamente inexistente en redes recurrentes sencillas, y a que los gradientes propagados a través de varios niveles o pasos temporales tienden a desaparecer con el tiempo, lo que se conoce como el desvanecimiento del gradiente. Para responder a este problema surgen las redes neuronales recurrentes LSTM y *Gated*

Recurrent Unit (GRU), cuya idea es crear caminos a través del tiempo cuyas derivadas no desaparezcan.

Las RNNs necesitan ser entradas de manera diferente a las redes convencionales, esto es porque, para el caso de las redes recurrentes se necesita propagar la información a través de las conexiones recurrentes entre pasos. Los algoritmos más comunes de entrenamiento para RNN son el llamado *Backpropagation Through Time* (*BPTT*) y *Real-time Recurrent Learning* (*RTRL*).

Por otro lado, las redes neuronales recurrentes o muy profundas son difíciles de entrenar, y a menudo, sufren además del problema de la explosion del gradiente. La explosion del gradiente en redes profundas de perceptrones multicapa, hace referencia a len el mejor de los casos, no puede aprender dos los gradientes explosivos pueden dar como resultado una red inestable que, en los datos de entrenamiento, y en el peor de los casos, genera valores de peso de *NaN - Not a Number* (puede ser entendido como valor no definido) que ya no se pueden actualizar.

El problema de la explosion del gradiente se refiere a un gran incremento en la norma del gradiente durante el entrenamiento. Tales eventos son debidos a la explocion de grandes. en el artículo publicado por Razvan Pascanu, Tomas Mikolov, Yoshua Bengio, (2013*)*.

El entrenamiento de las redes neuronales recurrentes (RNN) parece simple en el sentido de que se disponen de un conjunto de matrices de pesos, sin embargo, es extremadamente complejo debido a sus conexiones recurrentes. Por ejemplo, a medida que se multiplican todas las matrices de pesos en propagación tenemos que hacer lo mismo en la propagación hacia atrás. A medida que avanzamos hacia atrás, la señal puede volverse demasiado fuerte o demasiado débil; este es el problema de explosión o desaparición del gradiente, respectivamente.

Las redes neuronales recurrentes pueden extraer dependencias temporales, por lo tanto, las mismas son utilizadas en aplicaciones tales como procesamiento de voz, control no-markoviano, análisis de series temporales, composiciones musicales. Las redes neuronales recurrentes deben aprender de las entradas anteriores que han sido almacenadas para producir una salida deseada. El algoritmo de la propagación hacia atrás, sin embargo, sufre de un tiempo de aprendizaje demasiado largo, por ejemplo, con los algoritmos BPTT "*backprop though time*" , RTRL *"Real time Recurrent Learning*" la señal de error que retrocede en el tiempo, tiende a desaparecer. Sepp Hochreiter, (2016).

3.3.7. Redes neuronales de memoria a largo y corto plazo (LSTM).

LSTM es un tipo de arquitectura de red neuronal RNN, usado en el campo de aprendizaje profundo. Los autores Sepp Hochreite, et.al (1997). proponen el modelo LSTM para superar el problema del desvanecimiento del gradiente y así poder aprender dependencias a largo plazo. En este modelo se reemplazan los nodos de la capa oculta por unos nodos especiales denominados memory. Una celda LSTM está compuesta por tres componentes principales: una puerta de entrada (*input gate*), una puerta de salida (*output gate*) y una puerta de olvido (*forget gate*). La puerta de olvido es esencial, ya que permite el reinicio del estado interno de la LSTM, facilitando la gestión de la memoria a largo plazo. Esta célula es capaz de retener valores durante intervalos de tiempo arbitrarios, mientras que las tres puertas controlan el flujo de información en la célula. En resumen, la arquitectura de una red LSTM incluye una serie de conexiones recurrentes dentro de subredes, conocidas como bloques de memoria, diseñadas para mantener estados a lo largo del tiempo y regular el flujo de información a través de unidades no lineales. Van Houdt, G. et al. (2020).

Las redes LSTM, a diferencia de las redes neuronales estándar, incorporan conexiones hacia atrás. Esto les permite procesar no solo puntos de datos aislados, sino también secuencias completas de datos, tales como el lenguaje o vídeos. Por ejemplo, las LSTMs se han utilizado en tareas de reconocimiento de voz (Sak et al., 2014). Además, se emplean en otros ámbitos, incluyendo el reconocimiento de escritura a mano y la detección de anomalías en tráficos de red Graves et al. (2009).

El estudio realizado por A. Pulver, et. al. (2017) muestra ilustración y resume el proceso de salida de LSTM, en lo cual se detalla lo siguiente: x_t es el vector de entrada en el tiempo t, y y_t es la salida de la capa de red en el tiempo t, $\sigma = \frac{1.0}{1.0 + e^{-x}}$. La formulación estándar de LSTM incluye una puerta de entrada, una puerta de salida y, generalmente, una puerta de olvido (forgot gate).

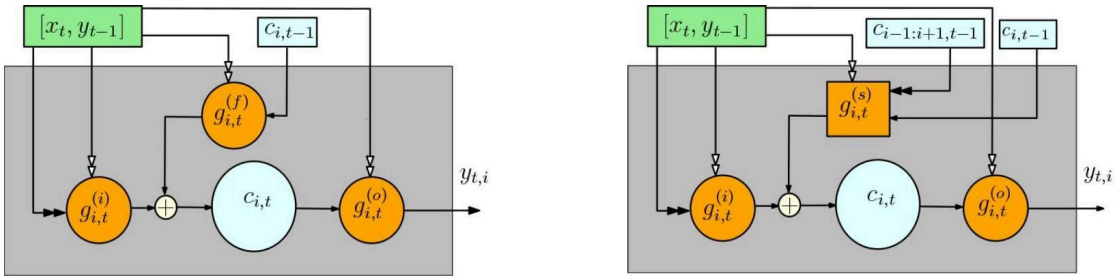


FIGURA 12. A. Pulver. et al. (2017).

3.3.8. Proceso *forward* (Cálculo de la salida en bloque LSTM).

Según Pulver et al. (2017), la salida se puede calcular de la siguiente manera:

$$a = f(W \cdot [x_t; y_{t-1}] + b) \quad (1)$$

$$g^{(i)} = \sigma(W_{g_i} \cdot [x_t; y_{t-1}] + b_{g_i}) \quad (2)$$

$$g^{(o)} = \sigma(W_{g_o} \cdot [x_t; y_{t-1}] + b_{g_o}) \quad (3)$$

$$g^{(f)} = \sigma(W_{g_s} \cdot [x_t; y_{t-1}] + b_{g_s}) \quad (4)$$

$$c_t = g^{(i)} \odot a + g^{(f)} \odot c_{t-1} \quad (5)$$

$$y_t = g^{(o)} \odot f(c_t) \quad (6)$$

Aquí se detalla cada componente de la ecuación, simplificando la comprensión de su funcionamiento en una red LSTM, defiendo cada elemento de la ecuación previa:

(1) **Activación Lineal**: Esta ecuación representa una activación lineal que resulta de aplicar una transformación lineal a la entrada, que es la combinación del estado oculto anterior y la entrada actual. Se utilizan los pesos W y los sesgos b, junto con una función de activación como la tangente hiperbólica ($tanh$) o la función de unidad lineal rectificada (ReLU).

(2) **Puerta de Entrada**: Esta parte de la ecuación define la puerta de entrada, que determina cuánta información nueva se permite incorporar al estado de la celda. Utiliza la función sigmoide σ, que limita los valores entre 0 y 1, para regular el flujo de información nueva.

(3) **Puerta de Salida**: Regula la cantidad de información del estado de la celda que se transfiere al estado oculto. Esta duplicidad en la redacción se debe a la importancia de la puerta de salida en el control del flujo de información hacia el estado oculto.

(4) **Puerta de olvido**: Esta ecuación describe la puerta de olvido, que decide cuánta información del estado anterior debe conservarse o eliminarse. Es crucial para permitir que la red olvide la información que ya no es relevante.

(5) **Actualización del Estado de la Celda**: El estado de la celda en el tiempo t Se actualiza con base en la nueva información filtrada por la puerta de entrada y la información anterior ajustada por la puerta de olvido.

(6) **Actualización del Estado Oculto**: El estado oculto en el tiempo t se calcula activando el estado de la celda mediante la función f y luego ponderando este resultado con la activación de la puerta de salida.

Finalmente, se explica brevemente cómo se entrena una LSTM utilizando el algoritmo de Backpropagation Through Time (BPTT), destacando la integración de estos componentes en el proceso de aprendizaje.

3.3.9. Entrenamiento y ajuste en LSTM - BPTT.

Según Paul y colaboradores (1990), el entrenamiento de redes neuronales recurrentes (RNN) utilizando el algoritmo de retropropagación a través del tiempo (BPTT) es una herramienta potente para una amplia gama de aplicaciones. Estas incluyen el reconocimiento de patrones, modelos dinámicos, análisis de sensibilidad, modelos econométricos, estructuras de lógica borrosa y modelos dinámicos de fluidos Paul et al, (1990). Las RNN que emplean unidades LSTM pueden ser entrenadas de manera supervisada en conjuntos de secuencias de entrenamiento, utilizando un algoritmo de optimización como el descenso de gradiente, combinado con BPTT para calcular los gradientes necesarios durante el proceso de optimización.

3.3.10. Algoritmo Backpropagation Through Time (BPTT).

El algoritmo Backpropagation Through Time (BPTT) es una extensión del método de retropropagación estándar utilizado en redes neuronales, adaptado para entrenar redes neuronales recurrentes (RNN) que procesan secuencias de datos. Este algoritmo gestiona la naturaleza secuencial de los datos al desplegar la red neuronal recurrente en el tiempo para cada instancia de la secuencia, formando así una red alimentada hacia adelante temporalmente extendida. Luego se aplican las técnicas tradicionales de retropropagación para calcular los gradientes y actualizar los pesos de la red en función de la diferencia entre la salida pronosticada y la real. Es fundamental en tareas como el procesamiento del lenguaje natural o el análisis de series temporales.

Finalmente, los parámetros que pueden ser ajustados en el entrenamiento, son detallados en la TABLA 7 del ANEXO del presente trabajo.

3.4. Métricas de evaluación de modelos.

3.4.1. El criterio de información de Akaike (AIC).

AIC es un estándar utilizado para medir y comparar la calidad de los modelos estadísticos. Su objetivo es seleccionar el modelo que mejor se ajuste a los datos, equilibrando entre la simplicidad del modelo y su capacidad para explicar la variabilidad de los datos. Este criterio es especialmente útil en situaciones donde se comparan múltiples modelos que no son anidados, es decir, modelos que no pueden ser simplificados uno dentro del otro mediante la restricción de parámetros. Es muy útil en nuestro caso de estudio, experimentalmente, nos permite seleccionar aquel modelo SARIMA que mejor ajuste nuestros datos. Es de interés mencionar que también se utiliza otra clase de parámetros como es BIC, que se verá más adelante.

El AIC se basa en la teoría de la información, la fórmula matemática del AIC es:

$$AIC = 2k - 2ln(L).$$

En la ecuación previa, k es el número de parámetros en el modelo estadístico, y L es el máximo valor de la función de verosimilitud para el modelo estimado.

3.4.2. Error cuadrático medio (*MSE*).

En el ámbito de la estadística, el error cuadrático medio (MSE, por sus siglas en inglés) es una medida que evalúa el promedio de los cuadrados de los errores. Esto implica que calcula la diferencia al cuadrado entre el valor estimado por un estimador y el valor real que se pretende estimar. El MSE actúa como una función de riesgo, reflejando el valor esperado del cuadrado de la desviación, también conocida como pérdida cuadrática. Esta discrepancia puede surgir tanto de la variabilidad inherente a los datos como de la posible omisión de información relevante por parte del estimador, que podría mejorar la exactitud de las estimaciones.

Además, el MSE se define como el segundo momento del error respecto al origen, lo que significa que abarca tanto la varianza del estimador como su sesgo. Para un estimador no sesgado, el MSE se corresponde con la varianza del estimador. Dado que mide el cuadrado

de las unidades de la magnitud estimada. *MSE* se expresa en las mismas unidades cuadradas de dicha magnitud. Si $\hat{Y}$ es un vector de N predicciones y Y es el vector de los verdaderos valores, entonces,

(el estimado) del predictor es $ECM = \frac{1}{n}\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$.

3.4.3. Raíz del error cuadrático medio *(RMSE).*

Por otro lado, la raíz cuadrada del ECM resulta en el error estándar de la media (RMSE o RMSD, por sus siglas en inglés), el cual se presenta en las mismas unidades que la magnitud estimada. Para un estimador no sesgado, el RMSE corresponde a la raíz cuadrada de la varianza, es decir, la desviación estándar.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}.$$

3.4.4. Bayesian information criterion (*BIC*).

Los modelos con un menor valor del *Bayesian Information Criterion* (BIC) son generalmente preferidos, ya que indican un mejor equilibrio entre la calidad del ajuste y la complejidad del modelo. El BIC se basa, en parte, en la función de verosimilitud, y penaliza más severamente los modelos con mayor número de parámetros, en comparación con otros criterios como el *Akaike Information Criterion* (AIC), con el cual guarda una estrecha relación (Schwarz, 1978).

Formalmente BIC es definido como:

$$BIC = kln(n) - 2ln(\hat{L}).$$

Los parámetros de la ecuación son los siguientes:

- $\hat{L}$ Es el máximo valor de la función de verosimilitud de un modelo M.
- n es el número de observaciones.
- k Es el número de parámetros estimados del modelo o equivalente, el tamaño muestral.

3.4.5. Hennan-Quinn Criterion (*HQIC*).

En el ámbito de la estadística, el criterio de información de HQ representa una herramienta valiosa para la selección de modelos SARIMA. Este criterio se ofrece como una alternativa eficaz a los métodos tradicionales de AIC y BIC. La fórmula para calcularlo es:

$$HQIC = nlog(\frac{RSS}{n}) + 2klog(log(n)).$$

en la ecuación anterior, k es el número de parámetros, n es el número de observaciones y RSS Es la suma residual de cuadrados que resulta de una regresión lineal u otro modelo estadístico.

3.3.6. Diagnóstico de los modelos.

La evaluación y diagnóstico de un modelo implica la verificación de que se cumplan las hipótesis fundamentales asociadas a los residuos. En el caso del modelo SARIMA, es crucial que estos residuos demuestran: una media marginal de cero, una varianza marginal constante, ausencia de autocorrelación a cualquier nivel de retardo, y una distribución normal. De los cuatro criterios para las distribuciones marginales, el primero es relativamente flexible, es decir, un modelo puede presentar errores significativos y, aún así, cumplir con esta condición. La tercera condición, la ausencia de autocorrelación para cualquier retardo, es esencial para garantizar que el modelo sea adecuado. En cuanto al modelo LSTM, es importante evaluar las curvas de predicción y asegurar que no exista sobreajuste, un fenómeno común en este tipo de algoritmos, utilizando las métricas previamente definidas de *MSE* y *RMSE*. ver en 3.4.2 y 3.4.3 respectivamente.
Dunning, T., et. al (2014). destacan que la detección de anomalías requiere modelar comportamientos normales para identificar los anormales.

3.5. Evolución del cultivo de soja en la región.

Describir la evolución de los cultivos de soja es fundamental para establecer una línea base de comportamiento esperado del cultivo en condiciones normales. Al tener esta referencia, es posible compararla con otros datos de cultivos que han sido afectados por diferentes tipos de disturbios, como enfermedades, plagas, o condiciones climáticas adversas. Según lo observado en la FIGURA 13.

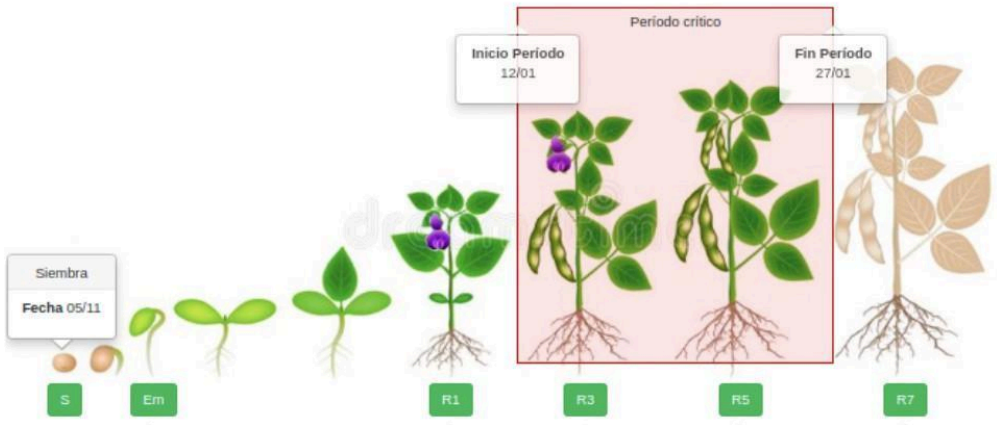


FIGURA 13. Evolución ideal de un cultivo de soja. Fuente: CONICET (2020).

LA FIGURA 13 describe la evolución del cultivo de soja desde su siembra hasta alcanzar su madurez completa. Idealmente, esta evolución se analiza a través de los estados fenológicos o las etapas del ciclo ontogénico del cultivo. A continuación, se definen algunos de los estados mencionados.

3.5.1. Emergencia (EM).

Esta etapa ocurre aproximadamente entre 5 a 10 días después de la siembra. La semilla germina y emerge del suelo. Es crucial que haya suficiente humedad en el suelo para que la semilla pueda germinar.

3.5.2. Floración (R1).

Este estado se alcanza aproximadamente de 35 a 65 días después de la siembra. La planta comienza a florecer, lo cual es señal de inicio de la reproducción. La primera flor suele aparecer en una de las nodrizas superiores de la planta.

3.5.3. Fructificación (R3).

Se observa alrededor de 20 a 30 días después de la R1. En esta fase, las vainas se están formando y llenando, lo cual es esencial para el rendimiento final del cultivo.

3.5.4. Inicio de llenado (R5).

Este estado se alcanza aproximadamente de 10 a 20 días después de la R3. Se caracteriza por el inicio del llenado de las semillas dentro de las vainas, momento en el cual la planta requiere una cantidad significativa de agua y nutrientes.

3.5.5. Madurez fisiológica (R7).

Se observa aproximadamente de 5 a 10 días después de la R6. Las semillas han alcanzado su peso seco máximo, y la planta comienza a perder color y marchitarse. La humedad de las semillas está alrededor del 50-60%.

3.6. herramientas para el análisis del estado vegetativo.

3.6.1. Índice NDVI.

Una de las definiciones más reconocidas y utilizadas para el Índice de Vegetación de Diferencia Normalizada (NDVI) puede atribuirse a Rouse et al. (1974) en su trabajo para el programa de satélites de recursos terrestres de la NASA.

El NDVI calcula la relación entre la diferencia en la absorción de la luz en el espectro del infrarrojo cercano (NIR) y la luz visible roja, y la suma de estas mismas bandas. Este índice ha probado ser extremadamente útil en la agricultura para monitorear el crecimiento estacional de las plantas y estimar su contenido de agua. También se utiliza para evaluar la respuesta de los cultivos a los nutrientes esenciales, las condiciones agronómicas, el potencial de rendimiento, el estrés vegetal y los impactos de las plagas, entre otros usos. Un ejemplo de su aplicación práctica fue publicado por Zhang Z. T. et al. (2014), donde se muestra cómo estos índices, en combinación con referencias agronómicas, ayudan a cuantificar diversos aspectos del desarrollo y la salud de los cultivos.

La fórmula para el cálculo del NDVI es:

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

El Índice de Vegetación de Diferencia Normalizada (NDVI) utiliza las longitudes de onda del infrarrojo cercano (NIR) y la luz roja visible (RED) para su cálculo. Según la fórmula, el NDVI se mide en un rango de valores entre -1 y 1. Un valor negativo generalmente indica la presencia de agua sobre la superficie terrestre, lo cual es atípico en áreas de

cultivo salvo en casos de inundaciones. Dado que las series representan la evolución de los cultivos de soja, se espera que el NDVI mantenga valores positivos, reflejando la actividad fotosintética de las plantas.

Las series NDVI muestran patrones complejos y no lineales, influenciados por una variedad de factores ambientales y antropogénicos. Estos incluyen variaciones estacionales, prácticas agrícolas y eventos climáticos extremos, todos los cuales pueden alterar significativamente la señal del NDVI y su interpretación en el monitoreo de la salud y el crecimiento de los cultivos.

3.6.2. Sentinel 2 (S2).

Los Sentinel son una nueva flota de satélites diseñada específicamente para proporcionar los abundantes datos e imágenes de que se nutre el programa Copernicus, de la Comisión Europea.

Este programa único de vigilancia medioambiental, está cambiando drásticamente la forma en que gestionamos nuestro entorno, entendemos y abordamos los efectos del cambio climático y protegemos nuestra vida cotidiana.

S2 lleva una innovadora cámara multiespectral de alta resolución, con 13 bandas espectrales que aportan una nueva perspectiva de la superficie terrestre y la vegetación. La combinación de la alta resolución y las nuevas capacidades espectrales, así como un campo de visión que abarca 290 grados.

3.6.3. Google Earth Engine (GEE).

En el presente trabajo, se hizo uso de GEE, específicamente para descarga, proceso y obtención del conjunto de datos números sobre el cual se trabaja.

GEE es una plataforma poderosa y versátil para el procesamiento y análisis de datos geoespaciales, especialmente útil para trabajar con imágenes satelitales, como las del Sentinel-2 (S2). Aquí te detallo algunas de sus utilidades principales en este contexto:

1. **Acceso a Gran Cantidad de Datos**: GEE proporciona acceso a una vasta biblioteca de imágenes satelitales, incluyendo las series completas del Sentinel-2. Esto permite a los usuarios obtener datos históricos y actuales sin la necesidad de almacenar grandes volúmenes de información localmente.
2. **Procesamiento en la Nube**: Al ser una plataforma basada en la nube, GEE realiza el procesamiento de datos en servidores remotos, lo que reduce significativamente

la carga computacional en los sistemas locales del usuario. Esto es especialmente útil para análisis a gran escala o para usuarios con limitaciones de hardware.

3. **Herramientas de Análisis Avanzado**: GEE incluye una amplia gama de algoritmos y herramientas preconstruidas para el procesamiento de imágenes, como la corrección atmosférica, la clasificación supervisada y no supervisada, y la detección de cambios. Estas herramientas son fácilmente accesibles y modificables a través de su interfaz de scripting.
4. **Escalabilidad**: La plataforma está diseñada para manejar operaciones a gran escala automáticamente. Esto significa que puede escalar los recursos según sea necesario para procesar grandes conjuntos de datos o ejecutar análisis complejos, facilitando así la gestión de proyectos extensos sin preocupaciones de rendimiento.

FIGURA 14. Ejemplo de imagen S2.

El Programa Copernicus es una iniciativa ambiciosa encabezada por la Comisión Europea en asociación con la Agencia Espacial Europea (ESA). Los Sentinel incluye imágenes de radar para todo clima de Sentinel-1A y Sentinel-1B, imágenes ópticas de alta resolución de Sentinel 2A y 2B, así como datos oceánicos y terrestres adecuados para el monitoreo ambiental y climático de Sentinel 3.

Para el presente trabajo, nos centraremos en el sensor Sentinel-2 MSI**,** el cual contiene un Instrumento multiespectral, el cual es suficiente para el cálculo necesario.

3.6.4. Python Tsaug y aumentación.

La aumentación permite crear versiones modificadas de los datos existentes, lo que efectivamente aumenta la cantidad de datos de entrenamiento disponibles. Esto es especialmente importante en series temporales, donde los patrones temporales pueden ser complejos y la cantidad limitada de datos puede impedir que el modelo aprenda estos patrones de manera efectiva.

Para la aumentación de series temporales, se ha seleccionado tsaug, una biblioteca en Python especialmente desarrollada para la edición y expansión de este tipo de datos. La librería proporciona una diversidad de quince métodos aplicables a las series temporales, ofreciendo una rica variedad de opciones para enriquecer los datos existentes y generar nuevas muestras que conserven las características fundamentales de la serie original. Dentro de esta gama de técnicas, resalta la implementación de aumentación por adición de ruido aleatorio, una estrategia que implica incorporar variaciones al dato original simulando fluctuaciones naturales y errores potenciales de medición. Este procedimiento busca fortalecer la capacidad de los modelos para adaptarse a variaciones auténticas en datos futuros, aportando al conjunto de datos instancias modificadas que, aunque distintas, reflejan patrones que podrían manifestarse en el fenómeno observado.

4. Metodología.

En esta sección, se detallan de manera ordenada los pasos metodológicos orientados a la detección de anomalías, con el fin de facilitar el logro de este objetivo. Se inicia discutiendo la herramienta GEE, de extrema utilidad para el preprocesamiento de las imágenes.

4.1. GEE y sus Aplicaciones.

Google Earth Engine (GEE) es una plataforma avanzada para el procesamiento de imágenes satelitales, que ofrece a los usuarios acceso a una extensa cantidad de datos geoespaciales y herramientas analíticas robustas. El uso de GEE para procesar imágenes del satélite Sentinel-2 (S2) presenta múltiples ventajas, tales como la automatización en el manejo de grandes conjuntos de datos, el acceso a herramientas de análisis avanzado, y la capacidad para realizar cálculos complejos, incluyendo la generación de series temporales de NDVI a gran escala.

A continuación, en 4.2, se enfatiza la importancia de GEE, destacando su integración con el monitoreo satelital, la investigación y los índices de interés.

4.2. Integración de NDVI en GEE para el Monitoreo Ambiental.

La integración de NDVI, descrita en la sección 3.6.1, con las funcionalidades de GEE, tratadas en la sección 3.6.2, resulta especialmente provechosa para monitorear cambios en la cobertura vegetal a lo largo del tiempo, evaluar la salud de los cultivos y gestionar los recursos naturales. Al definir y aplicar NDVI dentro de GEE, los investigadores y gestores de recursos optimizan la eficiencia en el manejo y análisis de datos, un aspecto crucial para la toma de decisiones informadas en sectores como la agricultura, la conservación, y la planificación urbana y rural.
Antes de proceder al ajuste de los modelos que se utilizarán en el estudio, se llevará a cabo el preprocesamiento de datos, detallado en la sección 4.3.

4.3. Preproceso de los datos.

4.3.1. La necesidad de la transformación de los datos.

El NDVI, un indicador esencial de la salud vegetativa que será analizado más adelante, se emplea ampliamente para monitorear la productividad y la dinámica de los cultivos a lo largo del tiempo. No obstante, el valor del NDVI y su evolución puede variar significativamente entre diferentes regiones debido a:

- **Nubes:** constituyen uno de los problemas más comunes en las imágenes satelitales. Cuando se presentan, es necesario aplicar técnicas que permitan suavizar la serie temporal —como la descrita en la sección 4.3.2, con el fin de reconstruir adecuadamente la evolución del cultivo analizado.
- **Diferencias en la cobertura vegetal en teledetección**: Algunas regiones pueden tener una vegetación naturalmente más densa o distintos tipos de cultivos que reflejan o absorben la luz de manera diferente, lo cual permite medir el vigor vegetal con índices que utilizan esta información.
- **Condiciones climáticas**: En regiones más áridas, el NDVI puede ser consistentemente más bajo debido a la escasez de agua, mientras que en áreas más húmedas podría observarse un NDVI más alto, esto, según los agrónomos, se

encuentran directamente vinculados a las condiciones de temperatura, precipitación y suelo necesarias para el desarrollo de ciertos cultivos. Para el caso de estudio planteado, las condiciones climáticas son ideales para el desarrollo de los cultivos de soja.

- **Prácticas agrícolas:** El manejo del suelo, el uso de fertilizantes y la rotación de cultivos pueden modificar la reflectividad de las plantas, afectando los valores de NDVI. Esto significa que la aplicación de fertilizantes o la rotación de cultivos pueden alterar los valores de NDVI. En estos casos, es recomendable realizar estudios que se apoyen en datos obtenidos, preferiblemente, directamente del campo.
- **La fecha de siembra:** Este factor crítico varía según la región y requiere ajustes específicos en los modelos agrícolas. Por esta razón, en este estudio se considera el tiempo transcurrido desde la siembra para el análisis, independientemente de si se dispone de la información en formato de fecha concreta.
- **Elección de la fecha de siembra, estacionalidad y clima**: La elección de la fecha de siembra depende en gran medida de las condiciones climáticas estacionales. Por ejemplo, en regiones templadas, las fechas de siembra están alineadas con las estaciones para evitar las heladas, mientras que en regiones tropicales, se pueden planificar en función de los patrones de lluvia, es importante la asesoría agronómica para esta clase de estudios en diferentes regiones. El NDVI es capaz de captar las características mencionadas previamente.
- **Duración del día**: Las variaciones en la longitud del día a lo largo del año afectan el crecimiento de los cultivos y, por tanto, impactan la fecha óptima de siembra.
- **Altitud y latitud**: Estos factores geográficos influyen en la temperatura y la exposición solar.

Como se muestra en la sección 4.3.1, múltiples variables influyen en la evolución del cultivo, y la elección de los modelos y sus ajustes depende en gran medida de la región de estudio, como bien lo mencionan los autores:. Malhotra et al. (2015)

4.3.2. Suavizado con Savitzky-Golay y medias móviles.

Detección de Nubes: Antes de aplicar el filtro, se utilizan máscaras de nubes, que son generadas automáticamente por los algoritmos de clasificación de imágenes satelitales.

Estas máscaras contienen bits específicos que indican la presencia o ausencia de nubes en cada píxel de la imagen.

Aplicación Condicional del Filtro: El filtro de Savitzky-Golay se aplica solo a los segmentos de la serie temporal donde la máscara indica la presencia de nubes. Esto permite suavizar los datos afectados por nubes, mientras que las áreas sin nubes se mantienen sin alteraciones, preservando la verdadera señal de NDVI donde las condiciones de observación son óptimas.

Conservación de la Calidad de Datos: Al utilizar el bit de las máscaras para guiar la aplicación del filtro, se conserva la integridad y la calidad de los datos originales en áreas despejadas. Esto es crucial para evitar la introducción de artefactos o distorsiones en la serie temporal de NDVI, asegurando que los análisis posteriores sean representativos de las verdaderas condiciones vegetativas.

4.3.3. Aumentación de las series temporales de NDVI.

En la fase de preprocesamiento, se adoptan procedimientos meticulosos que incluyen la descarga y cálculo del NDVI, la validación con observaciones directas de campo, y el manejo de artefactos como nubes a través de técnicas de interpolación y enmascaramiento para asegurar la integridad de los datos. Además, para simular la variabilidad natural y aumentar la robustez de los modelos, se lleva a cabo una aumentación de datos que introduce variaciones controladas y aleatorias.

4.3.4. Aumentación de las series temporales de NDVI.

En el caso de las redes LSTM, que están diseñadas específicamente para el análisis de secuencias, destacan por su habilidad para capturar dependencias a largo plazo. Esta característica las hace idóneas para ser empleadas en el análisis de múltiples series temporales, en contraste con métodos como SARIMA, que suelen aplicarse a una única serie. En cuanto a SARIMA, la aumentación para modelos de este tipo generalmente se centra en extender la serie o en ajustar sus componentes estacionales y de tendencia, con el objetivo de preservar las características esenciales del proceso temporal. Para realizar este proceso, se empleo la herramienta tsaug. ver apartado 3.6.4.

4.4. División del conjunto de datos.

Se exploran dos modelos principales de análisis: el modelo SARIMA, que facilita la identificación de patrones estacionales y la configuración de parámetros específicos, y un modelo LSTM configurado en Keras. Este último se optimiza ajustando el número de neuronas, el tipo de optimizador, y la tasa de dropout, entre otros hiper-parámetros. La división de datos para el entrenamiento de los modelos se establece en un 80% para entrenamiento y 20% para pruebas, asegurando una evaluación adecuada del rendimiento del modelo.

4.5. Evaluación y Ajuste de los modelos.

Es crucial evaluar la efectividad del modelo en la detección de anomalías y realizar los ajustes necesarios para mejorar su precisión. Esto puede incluir la recalibración de parámetros, el cambio del tipo de modelo o la revisión del método utilizado para definir y calcular las desviaciones. Esto implica evaluar indicadores como el AIC, BIC, HQIC, MSE y RMSE, definidos en las secciones 3.4.1, 3.4.4, 3.4.2 y 3.4.3, respectivamente. Además, se llevan a cabo validaciones gráficas. Los modelos ajustados trabajarán con la división de datos definida en la sección 4.4.

4.5. Detección de anomalías comparación pronóstico y observado.

En este trabajo se emplea el método estadístico Z-Score para la detección de anomalías. Este método evalúa cada punto de datos y clasifica como anomalías aquellos cuyos Z-Scores superen un umbral preestablecido Shehu, A., et al., (2023). Por otro lado, el método para comparar los valores pronosticados con los observados se detalla en los siguientes pasos:

1. Obtener la evolución del cultivo desde la fecha de siembra $t_0 = 0$ hasta la fecha actual (días después de la siembra). Esto implica recopilar datos de NDVI desde la siembra hasta la fecha actual para el lote 09.
2. Comparar cada punto pronosticado con el valor observado de NDVI en la serie temporal. Calcular la diferencia entre el valor pronosticado y el valor observado

para cada punto en el rango [0,dds], donde dds representa los días después de la siembra.

3. Por cada punto pronóstico, encontrar la diferencia entre el mismo y lo observado en la serie NDVI bajo estudio.
4. Se marcan anomalías a aquellos puntos como anomalías aquellos puntos que se encuentran fuera del rango $\left| NDVI_{pred} - NDVI_{obs} \right| \leq 3\sigma$. Cabe mencionar que el i-ésimo error se computa como $NDVI_{i\,pred} - NDVI_{i\;obs} = \varepsilon_i$, y que la distribución de ε es normal según lo visto observado en los experimentos, por lo cual pueden identificarse las anomalías según $NDVI_{pred} - NDVI_{obs} \leq 3\sigma$. Cabe mencionar, que la diferencia de $[-3\sigma, 3\sigma]$, cuyo acuerdo previo con expertos, aunque se validan también para el caso de SARIMA el caso de $\left| NDVI_{pred} - NDVI_{obs} \right| \leq 2\sigma$.

4.5. Aplicación de método de detección de anomalías en series temporales.

En el presente trabajo, se establece una metodología para detectar anomalías mediante la modelización de la evolución típica de una curva y su posterior uso para pronosticar valores futuros y detectar desviaciones significativas. Este proceso se desarrolla en varios pasos esenciales, descritos a continuación:

1. **Modelado de la Curva Típica:** Este proceso comienza con la modelación de la curva típica a partir de los datos históricos, utilizando los modelos seleccionados para el estudio. Este paso es crucial para comprender el comportamiento estándar de los datos y establecer una base para futuras comparaciones. Este enfoque se alinea con lo descrito por Dunning, T., et al. (2014), y requiere del preproceso de datos descrito en 4.3.1.
2. **Ajuste de los modelos y evaluación.**
3. **Pronóstico de Valores Futuros:** Una vez que se dispone de un modelo ajustado que reproduce fielmente los datos históricos, se emplea para pronosticar valores futuros. Este pronóstico asume que los patrones observados en el pasado continuarán en el futuro, permitiendo anticipar comportamientos típicos.

4. **Detección de Desviaciones (Anomalías):** La detección de anomalías se realiza comparando los valores observados en nuevos lotes de datos con los valores pronosticados por el modelo. Las desviaciones que superen un umbral predeterminado, definido en colaboración con ingenieros agrónomos y basado en la variabilidad típica de los datos, se clasifican como anomalías. Aplicando el método detallado en 4.5.

5. Aplicaciones.

En este estudio, se modela la evolución típica utilizando GEE y NDVI, discutidos en la sección 4.2, y descritos en la sección 4.1. El preprocesamiento de los datos se realiza conforme a la metodología establecida en la sección 4.3, específicamente para el cultivo de soja. Se ha desarrollado una curva ideal, detallada en la sección 4.5, que simula el proceso biológico de desarrollo de este cultivo. Mediante técnicas de aumentación de datos, se ajusta esta curva para reflejar de manera más precisa la evolución de los cultivos en la región estudiada. La curva modificada se ilustra en la FIGURA 15.

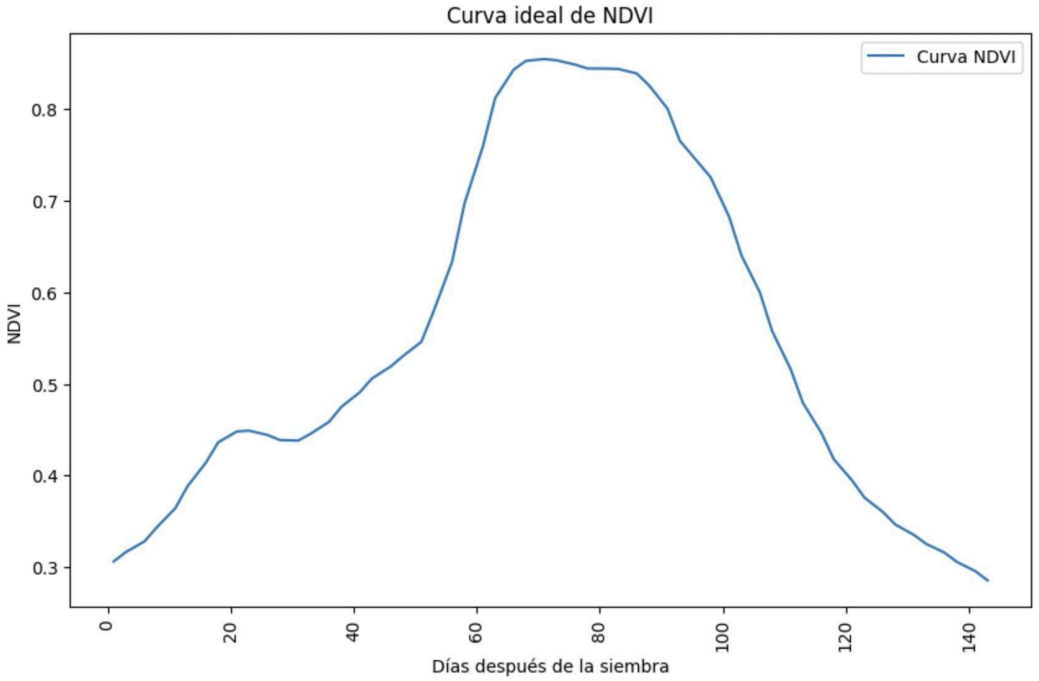


FIGURA 15. Curva ideal de NDVI.

En esta serie temporal, se analiza la evolución del NDVI de un cultivo de soja ubicado en el norte de la provincia de Santa Fe. Es importante destacar que, en vez de utilizar fechas específicas, se emplean los días transcurridos desde la siembra, lo que permite generalizar el análisis sin depender de una fecha de inicio exacta.

Al utilizar la fecha de siembra como referencia, es posible estimar el rango de NDVI esperado para un cultivo en cualquier día dado, de acuerdo con un patrón de crecimiento previsto. Esto se convierte en una herramienta útil para supervisar y gestionar la salud y el desarrollo del cultivo de soja durante su ciclo de vida.

Para perfeccionar los modelos y ampliar la base de datos, se expande la serie temporal que refleja este patrón de evolución. Este proceso involucra la introducción de errores aleatorios distribuidos de manera uniforme. En concordancia con lo discutido en el apartado 4.3.4 del presente trabajo.

5.1.1. Aumentación del conjunto de datos y el ajuste del modelo SARIMA.

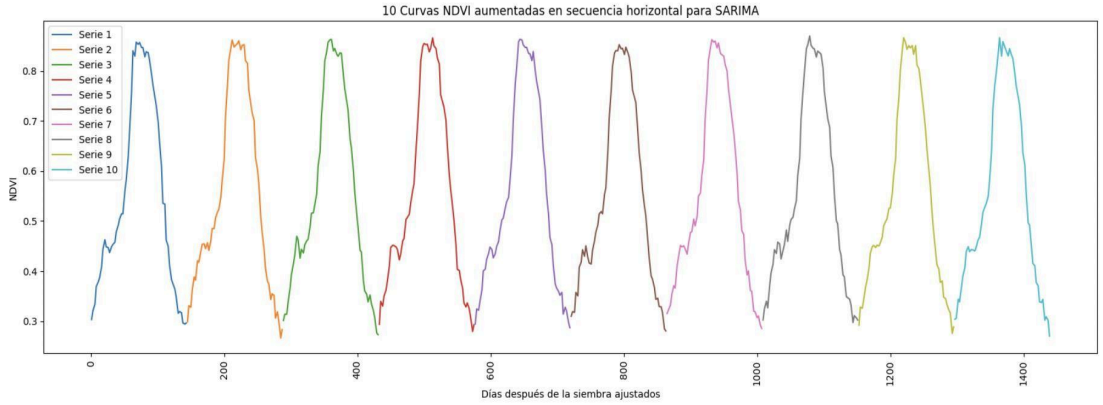


FIGURA 16. Aumentación de NDVI para SARIMA.

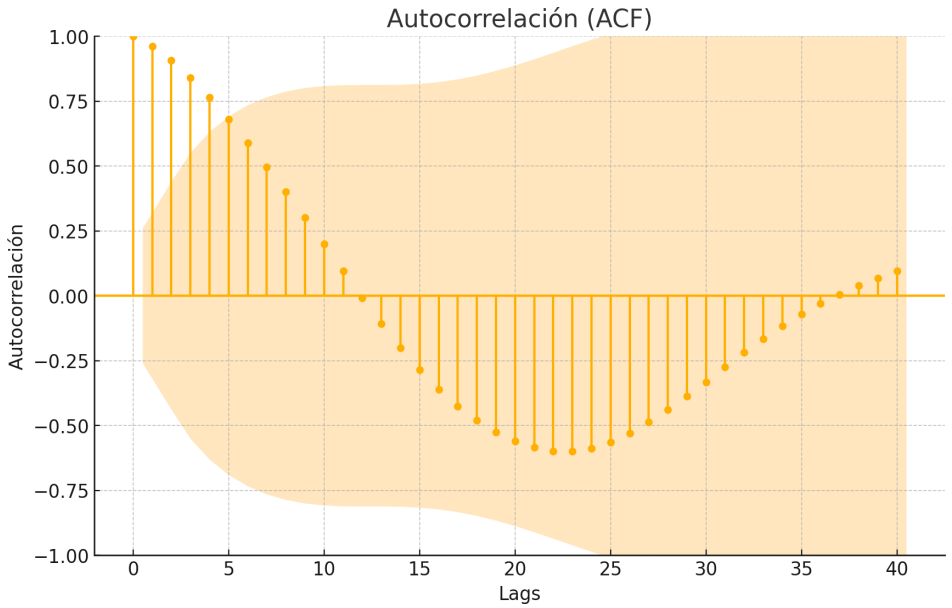


FIGURA 17. Función ACF.

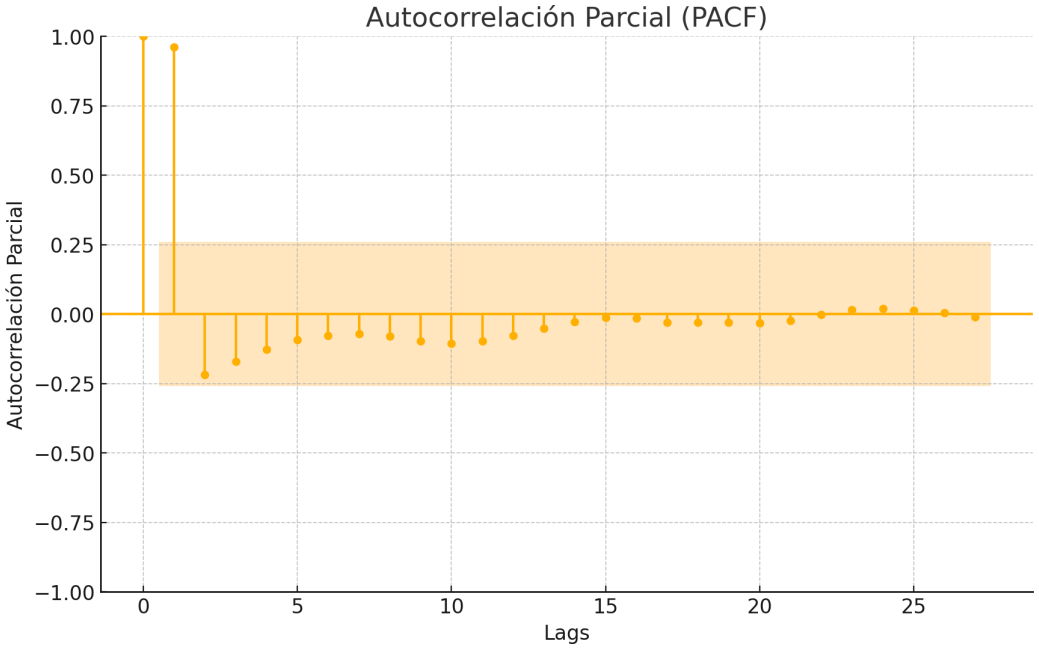


FIGURA 18. Función PACF.

Según lo observado en las FIGURAS 17 y 18, se pueden extraer conclusiones importantes respecto a las funciones de autocorrelación y autocorrelación parcial:

- **Autocorrelación (ACF)**, descrita en la sección 3.1.5, muestra cómo el NDVI se correlaciona con sus valores anteriores. Un decaimiento gradual observado en la

auto-correlación sugiere la posible presencia de un componente autorregresivo en la serie temporal.

- **Autocorrelación Parcial (PACF)**, definida en la sección 3.1.6, revela que los picos significativos en los primeros rezagos pueden ser puntos adecuados para establecer el parámetro de un componente autorregresivo (AR), tal como se discute en la sección 3.1.7 y su aplicación en un modelo SARIMA.

Teniendo en cuenta estas observaciones, se procede a realizar la aumentación de la serie en el apartado 5.1.2.

5.1.2. Aumentación del conjunto de datos para el ajuste de LSTM.

A partir de lo ya explicado en 4.3.4, se presenta la aumentación de la serie de la FIGURA 19.

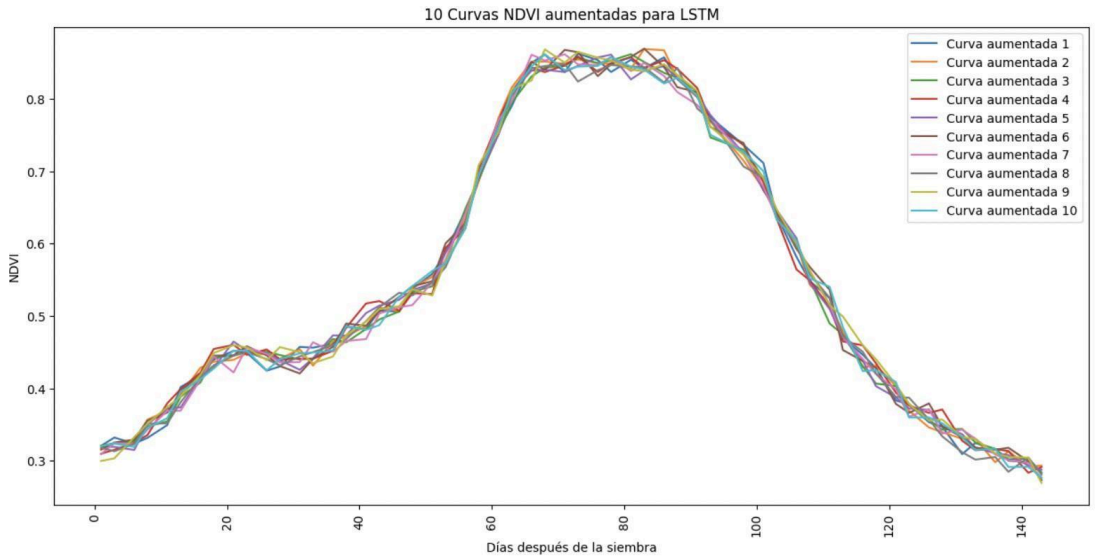


FIGURA 19 . Curva NDVI Aumentada para ajustar LSTM.

Se procederá al ajuste de los modelos SARIMA y LSTM, destacando que las aumentaciones se realizaron diez veces (n=10). Una vez ajustados estos modelos, se extraerán y presentarán conclusiones sobre cada uno de ellos. Es crucial para el lector comprender las diferencias en la aumentación aplicada a cada modelo, ya que los resultados de los ajustes son notablemente distintos.

Posteriormente, se presentarán los detalles de los ajustes de ambos modelos y se evaluarán, apoyándose en la discusión del apartado 4.5.

5.2. Ajuste de los modelos.

5.2.1. Ajuste del modelo SARIMA.

Se seleccionaron tres modelos distintos de redes neuronales LSTM luego de realizar una serie de experimentos variados. Siguiendo la metodología detallada en las secciones 4.4 (división de datos) y 4.5 (ajuste de modelos), y considerando la aumentación presentada en la sección 5.1.1, se ajustaron diversos modelos. De ellos, se eligieron cuatro para su evaluación, los cuales se presentan en un ranking con los cuatro mejores en la TABLA 3.

Modelo	AIC	BIC	HQIC
SARIMA(1, 1, 1)x(1, 1, 1, 6)	-13382.936	-13355.450	-13372.791
SARIMA(1, 1, 0)x(1, 1, 0, 6)	-12861.817	-12845.326	-12855.730
SARIMA(0, 1, 1)x(0, 1, 1, 6)	-13366.999	-13350.507	-13360.911
SARIMA(1, 0, 1)x(1, 0, 1, 6)	-13385.150	-13357.645	-13374.999

TABLA 3. AIC, BIC, HQIC de los modelos ajustados.

Los resultados de la tabla con los modelos SARIMA sugieren varias observaciones importantes para la selección del mejor ajuste para la serie temporal de NDVI:

1. **Comparación de Modelos:**
    - El modelo $SARIMA(1,0,1)x(1,0,1)_6$ muestra los valores más bajos de AIC, BIC y HQIC lo que indica que proporciona el mejor ajuste al conjunto de datos con la menor penalización por la cantidad de parámetros usados. Esto sugiere que es el modelo más eficaz en términos de balance entre el ajuste y la complejidad del modelo.
    - Los modelos $SARIMA(1,1,1)x(1,1,1)_6$ y $SARIMA(0,1,1)x(0,1,1)_6$ muestran buenos valores de ajuste, pero son ligeramente superiores (peores) en comparación con el modelo líder.
2. **Selección del modelo.**

- El modelo $SARIMA(1,0,1)x(1,0,1)_6$ no solo ajusta mejor los datos históricos, sino que su estructura implica que es suficientemente complejo para capturar la dinámica de la serie, pero no tanto como para ser sobreajustado. Esto lo hace ideal para proyecciones y análisis continuo. Podemos proceder a utilizar este modelo para pronósticos futuros o para estudiar más a fondo las características subyacentes de la serie de NDVI.

5.1.2. Visualización de residuos y predicciones.

Se generaron gráficos de residuos comparativos que confrontan los datos observados con las predicciones realizadas por el modelo $SARIMA(1, 0, 1)x(1, 0, 1)_6$, utilizando los días después de la siembra en el eje x. Esta elección permite una interpretación más intuitiva de la evolución temporal de la vegetación según el modelo.

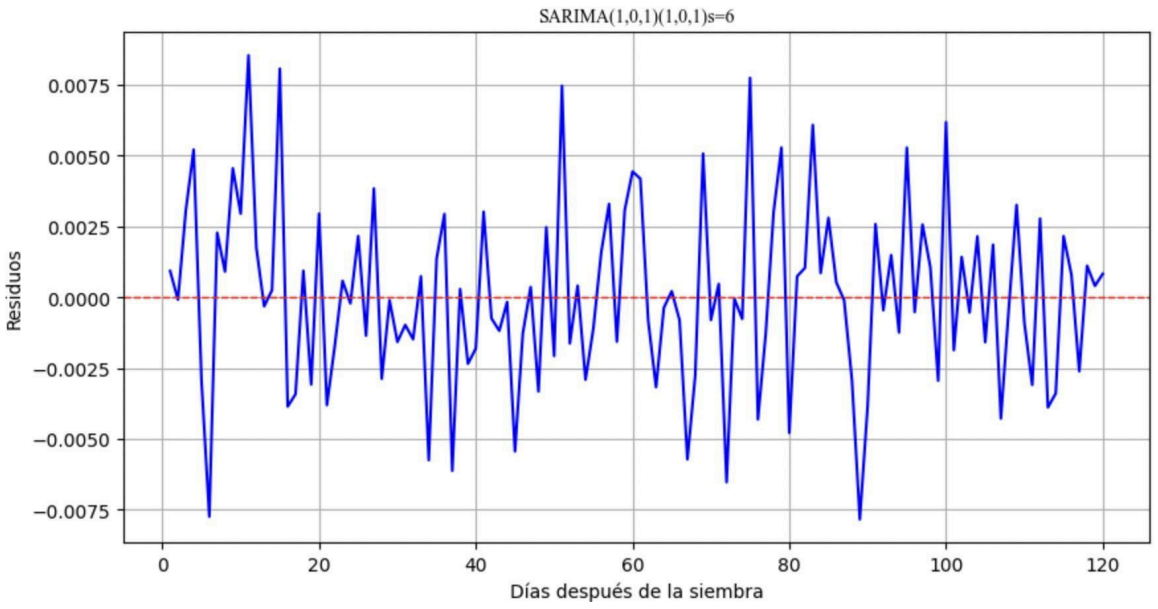


FIGURA 20. Residuos del modelo.

Conforme se observa en la FIGURA 20, se aplicó la prueba estadística de Ljung-Box a los residuos del modelo. Los resultados obtenidos son los siguientes:

- Estadístico de prueba (lb_stat): 8.900304
- P-value (lb_pvalue): 0.541591

La prueba de Ljung-Box se emplea para determinar si los residuos de una serie temporal son independientes, es decir, para verificar si constituyen un "ruido blanco", ver apartado

3.1.4. Un p-value superior a un umbral convencionalmente aceptado, $\alpha = 0.05$, indica que no existen evidencias suficientes para rechazar la hipótesis nula. Esto implica que los residuos analizados no presentan auto-correlaciones significativas, sugiriendo que el modelo ha capturado adecuadamente la estructura de dependencia en los datos, dejando únicamente fluctuaciones aleatorias que no están correlacionadas temporalmente.

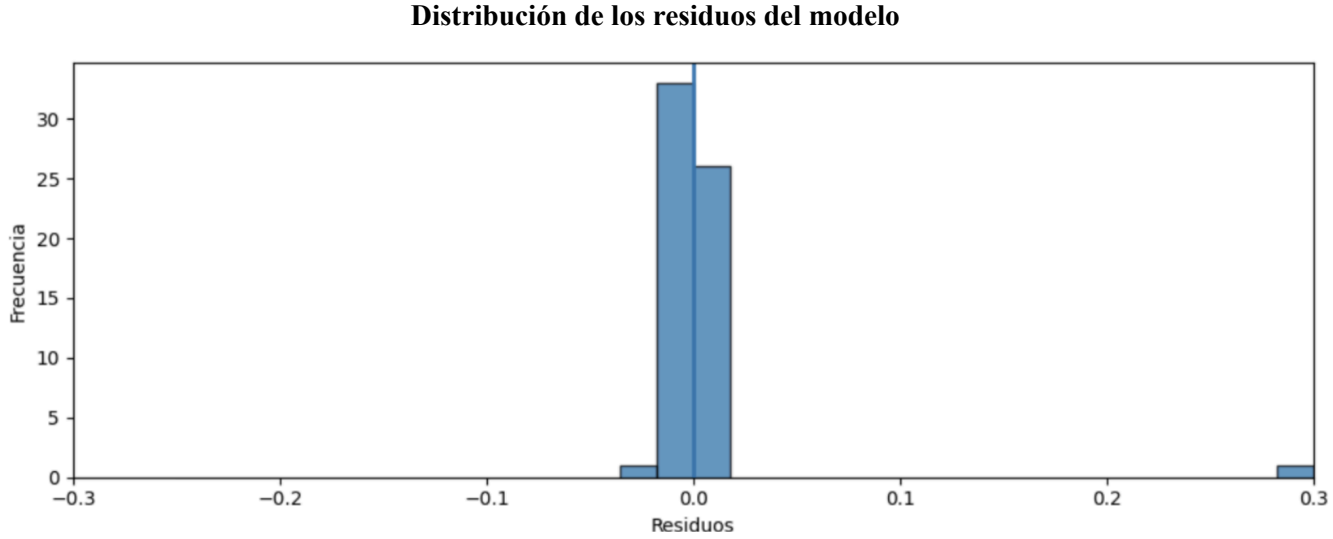


FIGURA 21. Distribución de los residuos del modelo.

La FIGURA 21 muestra un claro desvío del comportamiento normal, probablemente debido a la presencia de valores atípicos. Además, la distribución de frecuencias de los residuos al inicio de la serie tampoco parece normal. En la prueba de Shapiro-Wilk aplicada a estos residuos, se obtuvo un p-value de 0.009, significativamente inferior al umbral estándar de 0.05. Este resultado conduce al rechazo de la hipótesis nula, que postula una distribución normal de los residuos, señalando una posible inadecuación del modelo con los datos.

5.2.3. Conclusión del ajuste de SARIMA.

Según los experimentos $SARIMA(1, 0, 1)x(1, 0, 1)$, con s=6 sugiere un buen ajuste.
para los datos de NDVI en función de los días después de la siembra. Las predicciones

Las pruebas realizadas se ajustan razonablemente bien a los datos observados en el conjunto de prueba, lo que indica que el modelo puede ser útil para prever futuros valores de NDVI bajo condiciones similares. El análisis de residuos no muestra signos claros de autocorrelación y heteroscedasticidad no manejada, lo que sugiere que el modelo capta bien la estructura subyacente de los datos.

El análisis realizado en 5.2.1 sugiere que el modelo $SARIMA(1, 0, 1)x(1, 0, 1), \mathrm{s}{=}6$ muestra un modelo razonable. Sin embargo, la ligera desviación de la normalidad observada en la FIGURA 21, y la presencia de outliers podrían explorarse más a fondo, posiblemente mediante la incorporación de términos adicionales en el modelo o utilizando técnicas de transformación de datos para mejorar la normalidad.

Estos resultados apoyan en gran medida el uso del modelo para pronósticos a corto plazo, aunque siempre es recomendable realizar un monitoreo continuo del rendimiento del modelo a medida que se disponga de nuevos datos.

A continuación, se realiza una predicción con el modelo seleccionado.

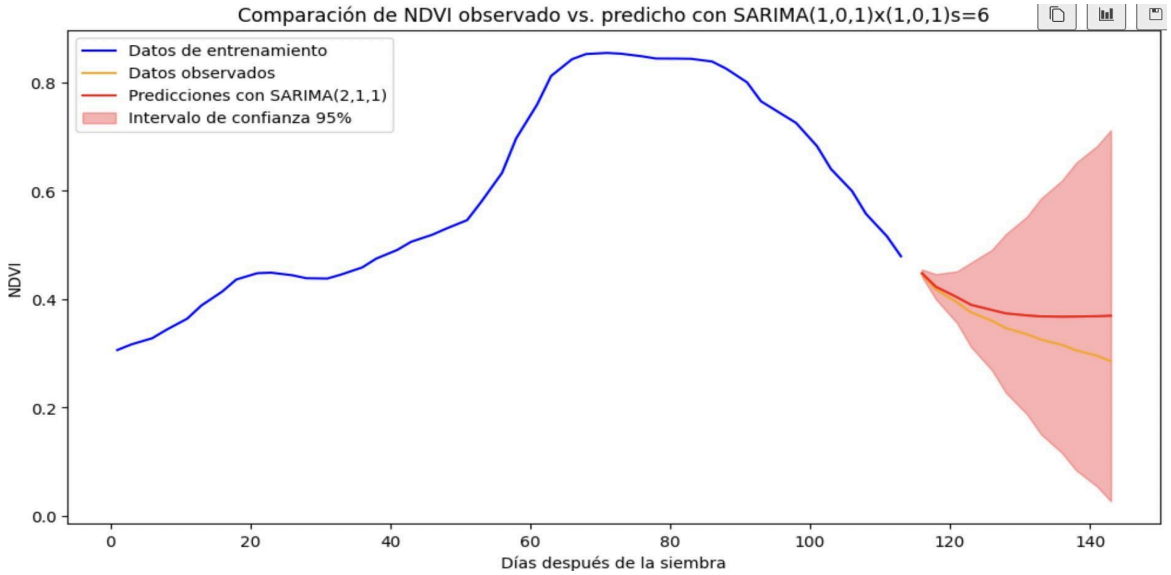


FIGURA 22. Predicciones con $SARIMA(1, 0, 1)x(1, 0, 1)_6$

A continuación, se procede a ajustar el modelo de LSTM, según el enfoque metodológico definido en el apartado 4.5.

5.2.2. Ajuste del modelo LSTM.

Se seleccionaron tres modelos distintos de redes neuronales LSTM luego de realizar una serie de experimentos variados. Como resultado, se estableció un ranking con los tres modelos de mejor desempeño, diferenciados principalmente por la configuración de sus unidades y optimizadores. Estos modelos están orientados a la predicción de valores futuros en series temporales, aplicados específicamente a datos escalados de NDVI. A continuación, se describen las principales diferencias entre ellos en términos de parámetros, lo que permite comprender cómo cada configuración incide en el rendimiento durante las etapas de entrenamiento y evaluación. El método de aprendizaje empleado fue el aprendizaje supervisado, como se detalla en la Sección 3.3.1, utilizando el algoritmo Backpropagation Through Time (BPTT) , según se explica en la Sección 3.3.10.

Nombre del modelo	Unidades LSTM	Optimizador	Tasa de aprendizaje
Modelo 1.	50	Adam	0.001
Modelo 2.	50	SGD	0.01
Modelo 3.	100	RMSprop	0.001

TABLA 4. Diferentes modelos ajustados LSTM.

A continuación, se presenta una imagen con los resultados de un experimento comparativo entre tres configuraciones de una red LSTM, cada una utilizando un optimizador distinto: Adam, SGD y RMSprop. El objetivo del experimento es modelar la serie temporal del NDVI. En cada uno de los gráficos se muestran las curvas de error correspondientes al proceso de entrenamiento y validación.

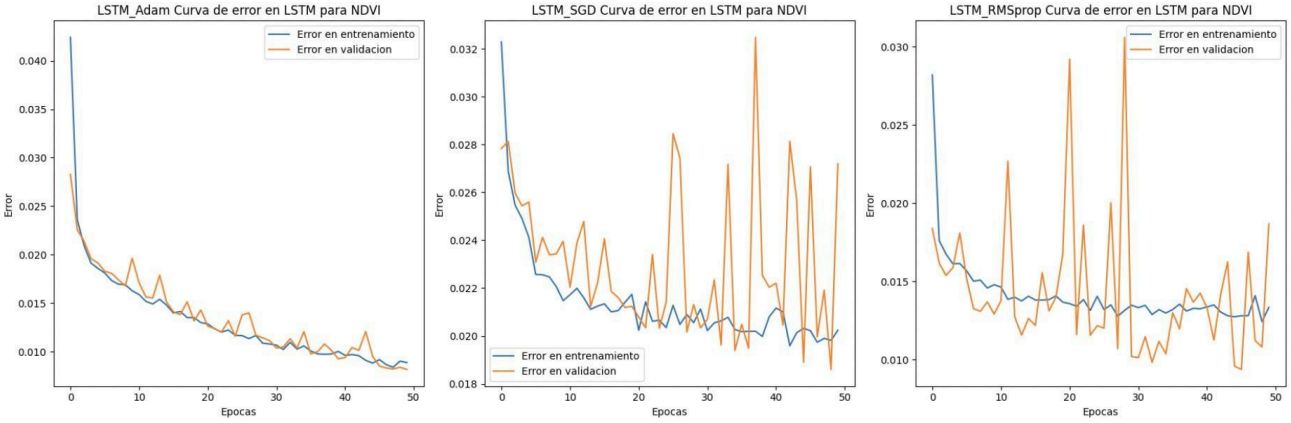


FIGURA 23. Curvas de resultado de testeo LSTM.

Interpretación de las Curvas.

Las curvas de error mostradas en los gráficos representan cómo el modelo mejora (reduce el error) a medida que aprende durante las épocas de entrenamiento. Una curva de error en entrenamiento descendente indica que el modelo está aprendiendo correctamente de los datos de entrenamiento, mientras que una curva similar en la validación sugiere que el modelo también generaliza bien a nuevos datos no vistos.

- **LSTM con Adam**: Muestra una disminución rápida entrenamiento como en validación, lo que indica un buen aprendizaje y generalización, aunque considerando la escala, se verá que SGD funciona mejor en el caso.
- **LSTM con SGD**: Las curvas son muy erráticas, especialmente en validación, indicando que el modelo puede estar aprendiendo, pero no generaliza bien, probablemente debido a una tasa de aprendizaje relativamente alta que causa oscilaciones en la convergencia.
- **LSTM con RMSprop**: Aunque muestra una disminución del error, las curvas son igualmente erráticas como con SGD, sugiriendo problemas similares de generalización.

Optimizador	MSE	RMSE
SGD	0.000743	0.02727
Adam	0.000771	0.02777
RMSprop	0.000803	0.02833

TABLA 5. Diferentes modelos ajustados LSTM.

El optimizador SGD tiene el menor MSE, métrica definida en 3.4.3, por lo tanto presenta la mejor capacidad de generalización entre los tres para este caso de predicción de NDVI con LSTM. Además, SGD utiliza una tasa de aprendizaje mayor que los demás (0.01).

5.2.3. Conclusión del ajuste de LSTM.

Basado en la estabilidad y la tendencia descendente de las curvas de error tanto en entrenamiento como en validación, el modelo LSTM con SGD - ver TABLA 5 - parece ser el más efectivo y eficiente para esta tarea específica. Exhibe un buen equilibrio entre aprendizaje y generalización sin las oscilaciones marcadas presentes en las otras dos configuraciones. Por tanto, para futuras tareas relacionadas con el modelado de NDVI mediante redes LSTM, se recomienda utilizar la configuración con el optimizador Adam. Para más detalles sobre los parámetros utilizados en las configuraciones del modelo, se puede consultar la TABLA 7 del ANEXO. Estos parámetros se mantuvieron como valores predeterminados durante los experimentos.

5.3. Aplicación de los modelos para la detección de anomalías.

En esta sección, se retoma el análisis de la curva del lote 09 ver (FIGURA 1), anteriormente identificado como dañado. Este lote se describió en detalle en la sección 3.3, donde se evidenció una anomalía de tipo puntual, conforme a lo observado en el apartado 3.2.2. Para abordar esta situación, se implementará una metodología de detección de anomalías que incluye el uso de dos modelos ajustados: inicialmente, un modelo SARIMA, seguido de un modelo LSTM. Posteriormente, se reflexionará sobre el desempeño de estos modelos y se realizará un análisis crítico de los resultados obtenidos. Este enfoque permitirá evaluar la eficacia de las técnicas utilizadas para detectar anomalías en este contexto específico.

La FIGURA 1 muestra claramente una anomalía que, según el análisis realizado en la sección 3.3.2 sobre estados fenológicos, sugiere que el cultivo se encuentra entre los estados R3 (Fructificación) y R5 (Inicio de llenado). Posteriormente, en la sección 5.3.1, se explora el modelo SARIMA aplicado al caso de estudio.

5.3.1. Detección con SARIMA.

Segun los experimentos detallados en 5.2.1 Ajuste de SARIMA, se realizan las predicciones, las prueba del modelo se realizaron con dos umbrales: 2σ y 3σ respectivamente, luego se realiza una descripción de lo observado.

Los experimentos indican que el modelo SARIMA con un umbral de 3σ no logra detectar anomalías en el lote. Sin embargo, al reducir el umbral a 2σ, el modelo es capaz de identificar variaciones más sutiles en los datos, las cuales podrían señalar potenciales anomalías. Este ajuste resulta especialmente útil para reconocer condiciones anómalas que, aunque no son extremadamente severas, pueden ser significativas para la salud de la vegetación. Las anomalías detectadas pueden incluir eventos menos intensos pero aún perceptibles, como cambios moderados en las prácticas de irrigación, variaciones en la fertilización, o efectos leves del clima que impactan el crecimiento vegetativo. Esta capacidad de detección más fina proporciona una herramienta valiosa para el monitoreo ambiental o agrícola, facilitando intervenciones más tempranas y ajustes en la gestión basados en la identificación precoz de posibles problemas.

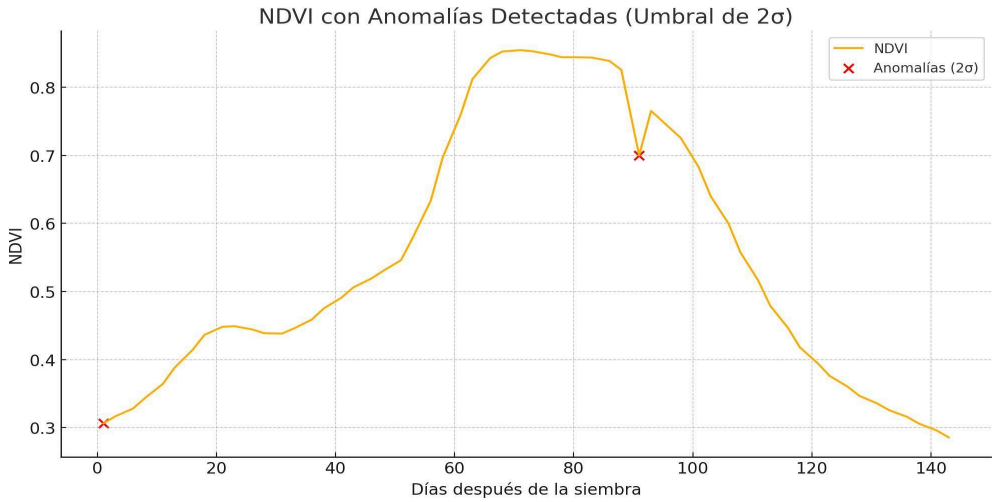


FIGURA 24. Anomalías con SARIMA.

Por otro lado, se realiza el mismo procedimiento con LSTM, bajo los resultados encontrados en 5.2.2. Ajuste del modelo LSTM.

5.3.2. Detección de anomalías con LSTM.

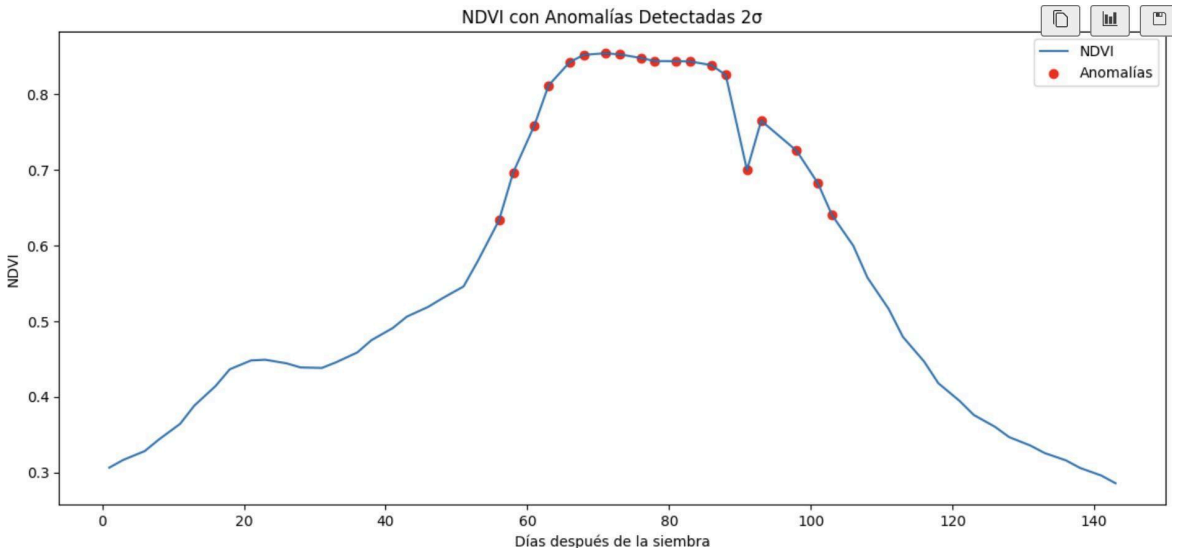


FIGURA 25. Anomalías con LSTM 2σ.

La FIGURA 25, muestra claramente en color rojo en formato de puntos, las anomalías detectadas en el cultivo de estudio para un umbral de 2σ, para el caso del LSTM.

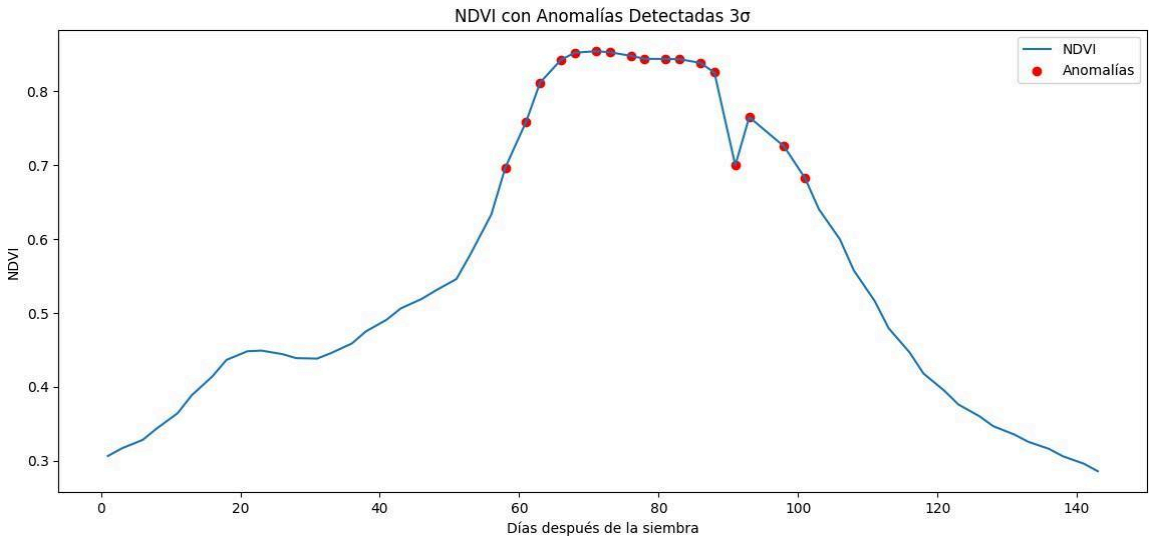


FIGURA 26. Anomalías con LSTM 3σ.

La FIGURA 26 ilustra claramente, mediante puntos rojos, las anomalías detectadas por el modelo LSTM en el cultivo estudiado, utilizando un umbral de detección de 3σ. Basándose en los experimentos realizados en las secciones 5.2.1 y 5.2.2, que abordan respectivamente los modelos SARIMA y LSTM, así como en las predicciones presentadas en las secciones 5.3.1 (SARIMA) y 5.3.2 (LSTM), y teniendo en cuenta los aspectos metodológicos desarrollados en el capítulo 4, se presenta a continuación la discusión de los resultados obtenidos.

Basándose en los experimentos realizados en las secciones 5.2.1 y 5.2.2, que abordan respectivamente los modelos SARIMA y LSTM, así como en las predicciones presentadas en las secciones 5.3.1 para SARIMA y 5.3.2 para LSTM, y teniendo en cuenta los aspectos metodológicos desarrollados en el capítulo 4, se presentan a continuación la discusión.

7. Discusión.

Los resultados obtenidos en este estudio demuestran, con sólido respaldo experimental y estadístico, que el modelo LSTM es más eficaz que el modelo SARIMA para la detección de anomalías en series temporales del índice NDVI en cultivos de soja del noreste de la provincia de Santa Fe.

Desde un enfoque cuantitativo, el modelo LSTM alcanzó un Error Cuadrático Medio (MSE) de $7{,}0 \times 10^{-4}$ y una Raíz del Error Cuadrático Medio (RMSE) de $2{,}7 \times 10^{-2}$ sobre el conjunto de prueba, valores que indican un excelente rendimiento predictivo. En contraste,

Aunque el modelo SARIMA presentó resultados favorables en términos de los criterios AIC, BIC y HQIC, mostró limitaciones relacionadas con la falta de normalidad en los residuos y otras dificultades que redujeron su eficacia, especialmente en pronósticos de largo plazo.

Estos resultados reflejan claramente una mejora significativa en la precisión predictiva del modelo LSTM, aspecto crucial en aplicaciones agronómicas, particularmente en la detección temprana de anomalías, donde incluso pequeñas variaciones en los valores del NDVI pueden implicar decisiones relevantes para la gestión y manejo de los cultivos.

Además, se destaca que el modelo LSTM fue capaz de identificar eventos atípicos superando el umbral de **3σ (tres desviaciones estándar)**, criterio estadístico comúnmente aceptado para la identificación de valores extremos. En contraste, el modelo SARIMA no logró detectar estos eventos, subestimando variaciones abruptas asociadas a condiciones ambientales adversas como heladas, tal como se evidencia en el análisis del Lote 09 (sección 5.3). Esto sugiere una mayor sensibilidad del modelo LSTM a patrones no lineales y rupturas en la estacionalidad típica del cultivo, lo cual es congruente con estudios previos como los de Hochreiter et al. (1997) y Pulver et al. (2017), que subraya la capacidad de las LSTM para modelar dependencias temporales de largo plazo.

Por otro lado, se identificaron valores de NDVI que superaron el umbral de **0.8**, los cuales podrían sugerir un sobreajuste del modelo o un falso positivo. Sin embargo, según observaciones de especialistas agrónomos consultados, tales valores son considerados inusuales para la región de estudio y no corresponden a fallas del modelo, sino a comportamientos atípicos atribuibles a condiciones atmosféricas, artefactos ópticos o errores de captura del sensor S2.

7. Conclusiones y futuras investigaciones.

7.1. Conclusiones.

Este estudio explora la detección de anomalías en cultivos agrícolas mediante el análisis de series temporales de NDVI, descritas en el apartado 3.6.2 y obtenidas del satélite Sentinel-2. Se evaluaron y compararon dos enfoques metodológicos principales: las redes neuronales de tipo LSTM y los modelos SARIMA, utilizando para ello imágenes satelitales del mencionado satélite. Comenzaremos validando las hipótesis:

7.1.1. Respuesta a las hipótesis.

Hipótesis 1: "Las series temporales de NDVI, obtenidas de cultivos de desarrollo típico mediante sensores remotos, son fundamentales para identificar anomalías en el crecimiento de los cultivos agrícolas de forma automática."
Respuesta: Esta hipótesis es validada por el estudio, ya que el uso de series temporales de NDVI permite observar y analizar las fluctuaciones en el crecimiento de los cultivos, detectando efectivamente anomalías que podrían ser indicativas de problemas ambientales o de gestión.

Hipótesis 2:
"Los modelos SARIMA y LSTM resultan extremadamente útiles para identificar patrones en la evolución del ciclo de cultivos en la región de interés, y pueden emplearse eficazmente para la detección de anomalías en estos procesos."
Respuesta: Según la investigación, ambos modelos muestran utilidad en identificar patrones y detectar anomalías, aunque hay diferencias en su efectividad dependiendo de la naturaleza específica de las anomalías y las características de los datos utilizados.

Hipótesis 3: "La utilidad de los modelos LSTM supera a los modelos SARIMA en la detección de anomalías en cultivos agrícolas, especialmente en la región estudiada.".
Respuesta: La tesis indica que, aunque ambos modelos son útiles, los modelos LSTM pueden superar a los SARIMA en ciertas condiciones debido a su capacidad para manejar dependencias de largo plazo en los datos, lo cual es especialmente relevante en el contexto de series temporales complejas y largas, como las que se encuentran en el seguimiento de cultivos.

7.1.2. Metodología.

El análisis se centró en la efectividad de estos modelos para identificar anomalías en el NDVI. Se abordaron diversos aspectos metodológicos, incluyendo el preprocesamiento de datos, la aplicación de técnicas de aprendizaje supervisado, y el entrenamiento y ajuste de los modelos. También se discutieron los fundamentos teóricos de las series temporales, la

estacionariedad, los componentes específicos de los modelos SARIMA y LSTM, y su aplicación práctica en la agricultura.

7.1.3. Efectividad de LSTM sobre SARIMA.

1. **Adaptabilidad a la No Linealidad**: LSTM es inherentemente adecuado para datos con patrones complejos y no lineales, crucial para analizar variables como el NDVI.
2. **Sensibilidad y Especificidad en la Detección de Anomalías**: LSTM demostró una mayor eficacia en captar puntos anómalos significativos, mientras que SARIMA requiere ajustes manuales para detectar anomalías relevantes, aumentando la tasa de falsos positivos.

Este estudio proporciona un marco robusto para futuras investigaciones y aplicaciones prácticas en la gestión agrícola mediante el uso de tecnologías avanzadas de monitoreo satelital.

7.2. Futuros trabajos de investigación.

7.2.1. Utilizar LLM para identificar anomalías.

La integración de Modelos de Lenguaje de Gran Tamaño (LLM) para la detección de anomalías en series temporales está marcando un hito significativo en la fusión de la inteligencia artificial con análisis de datos predictivos. Los LLMs, al aprovechar su avanzada capacidad de procesamiento del lenguaje natural, ofrecen una oportunidad sin precedentes para enriquecer el análisis predictivo y la detección de anomalías en campos tan diversos como las finanzas, la ciberseguridad y la salud. Estos modelos pueden complementar los enfoques cuantitativos tradicionales, proporcionando una profundidad analítica y precisión mejoradas al tratar con patrones complejos y matices en datos que los modelos tradicionales podrían pasar por alto. Su, J., et al. (2024).

7.2.2. Detección de fechas de siembra.

Este estudio se basa en la premisa de disponer de información referente a la fecha de siembra de los cultivos, aunque cabe destacar que la obtención de este dato puede presentar dificultades. Por ende, investigaciones futuras podrían dirigirse hacia la creación

de metodologías para estimar dicha fecha de siembra, lo que representaría un valioso complemento a esta tesis.

7.2.3. Uso de Autoencoders.

Numerosos estudios recientes han implementado algoritmos denominados AutoEncoders en la detección de anomalías. Una vía de investigación emergente podría explorar la comparación entre el rendimiento de modelos SARIMA y LSTM frente a los AutoEncoders, con el objetivo de identificar el más adecuado para el caso de estudio específico. Adicionalmente, se pretende determinar si la integración de AutoEncoders constituye una mejora significativa en comparación con las metodologías actuales

7.2.4. Otras fuentes de datos.

Este estudio ha empleado datos del satélite Sentinel 2; sin embargo, la utilización de otros sensores podría ofrecer resultados variados y, posiblemente, una mayor precisión. Se sugiere la exploración de imágenes de la empresa Planet o de la serie Landsat para el estudio de anomalías, lo que podría enriquecer los hallazgos. Además, evidencia de investigaciones relacionadas resalta el valor de los AutoEncoders, lo cual podría resultar en una mejora significativa de la precisión en los modelos LSTM.

7.2.5. Otros modelos de series temporales.

Además de los enfoques ya mencionados, existen modelos avanzados para la predicción de series temporales tales como XGBoost, LightGBM y los Modelos de Suavizado Exponencial en Espacio de Estados (ETS), cuya utilidad en la detección de anomalías podría ser evaluada en investigaciones futuras. Se sugiere también la exploración de la técnica de time warping para la detección de anomalías. Dicha técnica ajusta dos secuencias temporales a un mismo marco de tiempo, compensando diferencias en velocidad o distorsiones temporales, lo que podría ser especialmente relevante para comparar cultivos sembrados en distintas fechas, ofreciendo una perspectiva innovadora en el estudio de anomalías.

7.2.6. Explorar otras variables y realizar los experimentos.

La incorporación de variables exógenas podría potenciar el rendimiento de ambos modelos. Resulta interesante explorar qué sucedería en una comparación similar si se agregan más variables además del NDVI al análisis.

8. Bibliografía

- Andrade, F. H., et al. (2017). *Los desafíos de la agricultura argentina* (1.ª ed.). Ediciones INTA.
- Atzberger, Clement. (2013). Correction: Atzberger, C. Advances in Remote Sensing of Agriculture: Context Description, Existing Operational Monitoring Systems and Major Information Needs. Remote Sens. 2013, 5, 949–981. RS. 5. 949-981. 10.3390/rs5020949.
- Bermejo, J. (2022). *Redes neuronales aplicadas al análisis de datos del sistema público de bicicletas de la Municipalidad de Rosario* (Tesis/Trabajo final). Universidad Nacional de Rosario, Facultad de Ciencias Económicas y Estadística.
- Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2020). *Una revisión sobre la detección de valores atípicos/anomalías en datos de series temporales* (arXiv:2002.04236). arXiv.
- Chatfield, C. (2004). *The analysis of time series: An introduction* (6th ed.). Chapman and Hall/CRC.
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58. *(Si tuviste "Banerjee, Chandola & Kumar"*
- CONICET. (2020, 21 de julio). *CRONOSOJA: el software que ayuda a planificar la siembra de la soja*. Recuperado de: https://www.conicet.gov.ar/88502-2/
- Dunning, T., & Friedman, E. (2014). *Practical machine learning: A new look at anomaly detection*. O'Reilly Media.
- Forkel, M., et al. (2013). Trend change detection in NDVI time series: Effects of inter-annual variability and methodology. *Remote Sensing, 5*(5), 2113–2144.
- Gutiérrez, E. M. (2019). *Aplicación de modelos de redes neuronales recurrentes a la predicción de emisiones contaminantes de autobuses urbanos*. Madrid, España.
- Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015). *Long short term memory networks for anomaly detection in time series*. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN).
- Meroni, M., Fasbender, D., Rembold, F., Atzberger, C., & Klisch, A. (2019). *Near real-time vegetation anomaly detection with MODIS NDVI: Timeliness vs. accuracy and effect of anomaly computation options*. Remote Sensing of Environment, 221, 508–521. https://doi.org/10.1016/j.rse.2018.11.041
- Hecheltjen, A., Thonfeld, F., & Menz, G. (2014). Recent advances in remote sensing change detection – A review. En I. Manakos & M. Braun (Eds.), *Land use and land cover mapping in Europe* (pp. 145–178). Springer Netherlands.
- 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
- Hj Mohd Rhymee, N. H., Ratnayake, U., Rahman, E., & Shams, S. (2023). *Application of normalized difference vegetation index in agriculture to estimate rice yield*. *AIP Conference Proceedings, 2643*. https://doi.org/10.1063/5.0115666

- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice*.
- Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. En *Third ERTS Symposium* (NASA SP-351 I, pp. 309–317).
- Shehu, A., Maly, F., & Prazak, P. (2023). Outlier detection in time-series receive signal strength observation using Z-score method with Sη, scale estimator for indoor localization. *Applied Sciences, 13*(3), 1442. https://doi.org/10.3390/app13031442

9. Anexo.

Para desarrollar el modelo, se emplea la biblioteca Keras de Python, la cual se encuentra actualmente en su versión 2.11.0. Los hiper parámetros configurables son los siguientes:

Nombre de parámetro	**Descripción**
units	Entero positivo, dimensionalidad del espacio de salida.
activation	Función de activación para usar. Por defecto se utilizan $tanh\ x$**.** Si se pasa el valor None (No definido), no se aplica ninguna función, es decir, se computa $a(x) = x$.
recurrent activation	Función de activación para usar el paso recurrente. Por defecto, se utiliza la función sigmoide. Similar a otros parámetros, si se configura el valor None(No definido) a este parámetro, entonces no se aplica ninguna función de activación, es decir, se computa $a(x) = x$.
use bias	Es de tipo booleano (Verdadero/Falso), permite definir si las capas de la red utilizan el vector de bias o no. Usualmente es conveniente utilizar el valor de bias, el cual en la mayoría de los casos ayuda a la red a converger.

kernel initializer	Este parámetro permite la inicialización para la matriz de pesos del kernel, se usa para realizar una transformación lineal de las entradas. Por defecto se utiliza una función llamada "glorot_uniform" la cual calcula muestras de una distribución uniforme dentro del rango $[-\ limit,\ limit]$ donde $limit = \sqrt{\frac{6}{(n_{input}+n_{output}}}$ con n_{input} número de unidades de entrada y n_{input} número de unidades de salida.
recurrent initilaizer	El parámetro permite la inicialización para la matriz de pesos para el recurrent kernel, el cual es usado para la transformación lineal del estado actual.
bias initializer	Este parámetro permite la inicialización del vector bias, por defecto se utiliza el vector nulo.
unit forgot bias	El valor del parámetro en cuestión es booleano (Verdadero/Falso), por defecto Verdadero. Si el valor es verdadero, agrega uno al bias de forgot gate.
kernel regularizer	El parámetro permite establecer la función de regularización aplicada a los pesos de la matriz de kernel. Por defecto, ninguna función es aplicada, es decir, se computa $a(x) = x$.
bias regularizer	Esta función de regularización se aplica al vector de bias. Por defecto, ninguna función es aplicada, es decir, se computa $a(x) = x$.
activity regularizer	El parámetro especifica la función de regularización aplicada a la salida de la de la capa. Por defecto, ninguna función es aplicada, es decir, se computa $a(x) = x$.

kernel constraint	El parámetro especifica la función de restricción aplicada a la matriz de pesos de kernel. por defecto, ninguna función es aplicada, es decir, se computa $a(x) = x$.

recurrent constraint	El parámetro especifica la función de restricción aplicada a la matriz de pesos de Recurrent Kernel. Por defecto, ninguna función es aplicada. , es decir, se computa $a(x) = x$.
bias	El parámetro especifica la función de restricción aplicada al vector de bias. Por defecto, ninguna función es aplicada, es decir, se computa $a(x) = x$.
dropout	El parámetro especifica el valor real entre 0 y 1, fracción de unidades que se desean eliminar de la transformación lineal de las entradas. Por defecto, el valor 0 es establecido.
recurrent output	El parámetro especifica un valor entre 0 y 1. Fracción de unidades a eliminar de la transformación lineal del estado recurrente. Por defecto, el valor 0 es establecido.
return sequences	El parámetro especifica un valor booleano. Si devolver la última salida. en la secuencia de salida, o la secuencia completa. Valor predeterminado: Falso.
return state	El parámetro especifica un valor booleano, el cual indica si devolver el último estado además de la salida. Valor predeterminado: Falso.
go backwards	El parámetro especifica un valor booleano (por defecto Falso). Si es Verdadero, procesa la secuencia de entrada hacia atrás y devuelve como resultado la secuencia invertida.
stateful	El parámetro especifica un valor booleano (por defecto Falso). Si es verdadero, el último estado de cada muestra en el índice i en un lote se usará como estado inicial para la muestra del mismo índice en el siguiente lote.

time major	El parámetro refiere al formato de las entradas y las salidas de los tensores[1]. Si el valor es verdadero, la forma de la entrada y la salida es $[rezagos, bloque, característica]$, mientras que en el caso de falso, el formato es $[rezagos, bloque, característica]$. Usar $time\_major = Verdadero$ resulta un poco más eficiente porque evita transponer al principio y al final de los cálculos de la RNN.

TABLA 6. Resumen de los parámetros posibles de configuración para el modelo LSTM.

A la hora de ejecutar el modelo, se encuentran los siguientes parámetros configurables.

Parámetro	**Descripción**
inputs	El parámetro referencia un tensor en formato tridimensional con la forma [batch[2], timesteps[3], feature[4]].
mask	El parámetro referencia un tensor en formato binario de la forma [batch, timesteps] indicando si dado un time step debe ser enmarcado (opcional, por defecto nulo).
training	El parámetro representa un valor booleano, indica si la capa debe comportarse en modo entrenamiento o en modo inferencia.
initial state	El parámetro hace referencia a la lista de los estados iniciales de los tensores. Por defecto, nula, lo que causa la creación de tensores inicializados en 0.

TABLA 7. Resumen de los parámetros más importantes con su descripción.

Los elementos presentes en la TABLA 7 constituyen los hiper parámetros de la red neuronal. Se denominan así porque no son parámetros cuyos valores se estimen directamente, sino que son variables que influyen en la cantidad de parámetros y en las estimaciones asociadas a ellos. Es importante señalar que, por lo general, no está claro

[1] En matemática, un tensor es un objeto algebraico que describe una relación multilineal entre conjuntos de objetos algebraicos relacionados con un espacio vectorial.
[2] Cantidad de muestras disponibles.
[3] Timesteps: Cantidad de memoria de la red, para el caso de una secuencia de 50 puntos, en este caso, sería 50.
[4] Número de características que se analizan en la serie, para caso serie univariada, sería 1.

cuáles deberían ser los valores óptimos para los hiper parámetros. El ajuste de una red neuronal suele implicar la experimentación con diferentes valores para los hiper parámetros con el fin de encontrar una combinación que resulte en un buen desempeño.

Es importante aclarar que aquellos hiper parámetros que no están especificados en la configuración del modelo en el presente trabajo toman sus valores por defecto según lo definido en la TABLA 7.