



ESPECIALIZACIÓN EN CIENCIA DE DATOS
TRABAJO FINAL INTEGRADOR

**“UTILIZACIÓN DE REDES NEURONALES
RECURRENTE EN LA PREDICCIÓN DE TENDENCIAS
DEL MERCADO DE HARINA DE SOJA”**

ALUMNO: JUAN IGNACIO ALONSO (ID 105155)

TUTOR: DRA. MARCELA RICCILLO

RESUMEN

La volatilidad de los precios internacionales de la Harina de Soja impacta de manera significativa la economía de distintas industrias, gobiernos y, finalmente, la población. Los modelos Auto Regresivos de Media Móvil (ARIMA) constituyen una de las herramientas de análisis de series de tiempo más utilizadas. Sin embargo, el advenimiento de nuevas tecnologías de análisis y procesamiento de datos difundieron nuevas técnicas aplicables al estudio de series de tiempo, siendo Las Redes Neuronales Recurrentes del tipo LSTM una de ellas. En el presente estudio se compara la performance relativa de modelos ARIMA y RNR LSTM en la predicción de tendencias de precios de Harina de Soja.

ABSTRACT

Volatility of International Soymeal prices has a deep impact over economic performance of different Industries, Governments, and, lastly, population in general. Auto Regressive Moving Average models (ARIMA) are one of the most spread tools used for Time Series analysis. However, advent of new analysis and data processing techniques made available a vast set of tools applicable to Time Series study, Recurrent Neural Networks LSTM type is one of them. In this paper, relative performance of ARIMA and RNR LSTM for Soymeal price trend prediction is compared.

.

ÍNDICE

1. INTRODUCCION	5
2. ESTADO DE LA CUESTIÓN	6
3. PLANTEAMIENTO DEL PROBLEMA	9
4. JUSTIFICACIÓN DEL ESTUDIO	10
5. ALCANCE Y LIMITACIONES	11
6. HIPÓTESIS	12
7. OBJETIVOS	13
8. METODOLOGÍA	14
8.1. TÉCNICAS DE MACHINE LEARNING	14
8.2. HERRAMIENTAS	14
8.2.1. MODELO AUTOREGRESIVO DE MEDIA MÓVIL (ARIMA)	14
8.2.2. RED NEURONAL RECURRENTE (LSTM)	15
8.2.3. ARQUITECTURA LSTM	17
9. EXPERIMENTACIÓN	18
9.1. DEFINICIÓN DE VARIABLES	18
9.2. OBTENCIÓN DE LOS DATOS	18
9.2.1. SERIE DE PRECIOS DE HARINA DE SOJA	18
9.2.2. POSICIÓN DE FONDS (MMF)	19
9.2.3. INDICES DE VEGETACIÓN	19
9.2.4. STOCKS DE HARINA, ACEITE, Y POROTO DE SOJA EN USA Y PRODUCCION DE SOJA EN ARGENTINA, BRASIL Y USA.	20
9.2.5. DATASET FINAL	21
9.3. ANÁLISIS EXPLORATORIO	22
9.3.1. ANÁLISIS DE CORRELACIÓN DE LOS DATOS	22
9.3.2. ANÁLISIS DE SERIE DE TIEMPO	23
9.4. CONJUNTO DE ENTRENAMIENTO Y TESTEO	27
9.5. MODELO ARIMA	28
9.6. MODELO LSTM	28
9.7. MEDIA MÓVIL 20 DÍAS	31
10. RESULTADOS	32
11. CONCLUSIONES	33
11.1. TRABAJOS FUTUROS	34
12. REFERENCIAS BIBLIOGRÁFICAS	35

1. INTRODUCCIÓN

La volatilidad de los precios internacionales de la Harina de Soja (referenciados a nivel mundial en el marco del Chicago Board of Trade -CBOT) tiene grandes implicancias para una gran diversidad de sectores económicos a nivel global (Timmermann, 2006), generando incertidumbre e incrementando la necesidad de administrar los riesgos asociados a esta variabilidad. En la actualidad se dispone de una amplia gama de fuentes globales de información que proveen datos relacionados con las variables que impactan en las cotizaciones de la Harina de Soja. Sin embargo, hasta el momento son escasos los modelos desarrollados que combinen esta significativa oferta de datos, con el fin de reducir la incertidumbre asociada.

El presente trabajo pretende desarrollar un modelo predictivo de la tendencia de los precios de la Harina de Soja, a través de la utilización de Redes Neuronales Recurrentes (RNR) y la inclusión de distintas variables de influencia sobre dichas cotizaciones. Al mismo tiempo, este estudio persigue el objetivo de demostrar la superioridad de estos algoritmos (RNR) sobre el enfoque tradicional aplicado en esta cuestión durante los últimos años, i.e. modelos Auto Regresivos Integrados de Media Móvil (ARIMA).

Finalmente, el modelo desarrollado debe ser considerado por los distintos actores del mercado como una herramienta adicional al momento de definir sus estrategias, y no como un algoritmo de decisión para operar diariamente la cotización internacional de la Harina de Soja.

2. ESTADO DE LA CUESTIÓN

La Harina de Soja (SBM por sus siglas en inglés referidas a Soybeans Meal) constituye unas de las principales fuentes de proteína vegetal a nivel global, ofreciendo un contenido proteico de 35 a 40% y un excelente perfil de aminoácidos esenciales para la nutrición animal (USDA, 2021). A pesar que su inclusión en las dietas animales se redujo durante los últimos años a expensas de proteínas sintéticas y subproductos de la molienda de Maíz, actualmente aún cubre una importante proporción en las raciones alimenticias de los sistemas de producción porcina y avícola (entre 30-32% de la ración según Borja & Medel, 1998).

La producción mundial de SBM (mediante la industrialización del grano de Soja) se concentra principalmente en Argentina, Brasil y los Estados Unidos quienes, en conjunto, son responsables del 46% del volumen producido anualmente (247.5 MMT; USDA, 2022). Adicionalmente, el 85% del comercio global de este producto es abastecido por estos países (67.8 MMT; USDA, 2022), por lo que cualquier factor que afecte de manera directa o indirecta su producción y consecuente oferta, repercute significativamente en sus cotizaciones. A modo de ejemplo, la sequía registrada en Argentina durante el ciclo 2017/18 (bajo los efectos del fenómeno La Niña) produjo un incremento semanal de 14% en las cotizaciones futuras de SBM en el Chicago Board of Trade (en base a datos históricos de CME [CME, s.f.]).

La volatilidad de los precios de la SBM resulta de gran relevancia para todos los eslabones que conforman las cadenas agroindustriales. En primer lugar, tiene un impacto directo en la rentabilidad de los productores primarios de granos, dado que sus ingresos al final de la campaña dependen, en gran medida, de la capacidad de pago de la industria. En segundo lugar, influye de manera sustancial en las industrias productoras de carne (principalmente la industria Avícola y Porcina) dado que representa aproximadamente el 60.2% del costo total de producción (Arrieta J., 2015). Como resultado de esto, las fluctuaciones en las cotizaciones de los productos cárnicos reflejadas en los puntos de venta a los consumidores finales

también afectan a la población general. Finalmente, la mencionada volatilidad en las cotizaciones resulta crucial en países como Argentina, en los cuales los ingresos fiscales provenientes de este producto conforman una importante proporción del total (el SBM aportó al fisco argentino en 2021 USD 3.652.000.000, en base a datos de la Dirección de Mercados Agropecuarios [Dirección de Mercados Agropecuarios, s.f.]).

En consecuencia, disponer de modelos predictivos que permitan administrar el riesgo asociado a las constantes fluctuaciones de precios resulta una herramienta de especial relevancia para todos los sectores involucrados.

Los esfuerzos e investigaciones previas realizadas en el marco de este campo se concentraron principalmente en el desarrollo de modelos econométricos (series de tiempo). En el trabajo de Bessler (1982) se implementaron modelos auto regresivos y de media móvil; Wang et al. (Wang & Bessler, 2004) utilizaron modelos de vectores auto regresivos; mientras que Yang et al. (2003) experimentaron con modelos de corrección del error de vectores, y una amplia gama de variaciones de los mismos. A la vez, son escasos los trabajos en los cuales se han incorporado variables adicionales exógenas a las series de precios para el posterior desarrollo de modelos predictores.

El advenimiento de nuevas tecnologías de análisis y procesamiento de datos, y la utilización y aplicación de diversas técnicas de Machine Learning han demostrado un gran potencial en la predicción de precios de un gran conjunto de commodities agrícolas. Por ejemplo, Khamis et al (2014), utilizaron redes neuronales (backpropagation neural networks) para predecir precios futuros de Trigo mientras que Lopes (2018) aplicó modelos estadísticos de Machine Learning para estimar los precios del Café en Brasil. Ribeiro et al (2020), comprobaron la utilidad del ensamble de modelos bagging (Random Forest) y boosting (Gradient Boosting Machine) en la predicción de precios de Soja y Trigo en el estado de Paraná, Brasil.

Investigaciones recientes han demostrado que las Redes Neuronales poseen un gran potencial en el desarrollo de modelos predictivos de series de precios económicos o financieros, que en general tienden a estar influenciadas por un gran número de variables (Xu & Zhang, 2022) con un alto grado de precisión.

El presente estudio se centrará en la comparación de dos modelos predictores de la tendencia de los precios futuros de SBM en el mercado de CBOT. El primero de ellos consistirá en un modelo Autoregresivo de Media Móvil (ARIMA), mientras que el segundo utilizará RNR, incorporando variables exógenas a las cotizaciones para el desarrollo del mismo.

3. PLANTEAMIENTO DEL PROBLEMA

La volatilidad de los precios internacionales de la harina de Soja tiene grandes implicancias sobre los actores de distintas cadenas agroindustriales, afectando sus costos y retornos, al tiempo que también afectan los balances fiscales de naciones que dependen fuertemente de éstos. Diversos esfuerzos han sido concentrados en el desarrollo de modelos predictivos que permitan atenuar el riesgo asociado a estas oscilaciones, estando la mayoría focalizados en análisis auto regresivos de series de tiempo.

Durante los últimos años, nuevas técnicas de Machine Learning fueron aplicadas en este campo con el objeto de mejorar los resultados predictivos. La utilización de Redes Neuronales Recurrentes (RNR) fue una de las técnicas más empleadas para alcanzar este propósito.

Sin embargo, existen factores fundamentales que explican una porción importante de las oscilaciones de precios (como por ejemplo variables climáticas, técnicas y económicas) que hasta el momento no han sido contempladas mediante el análisis exclusivo de series de tiempo.

4. JUSTIFICACIÓN DEL ESTUDIO

El análisis de los factores exógenos o fundamentales que impactan en la variación de los precios internacionales (en especial su jerarquización y cuantificación) resultaría de información valiosa para amplios eslabones de la cadena agroindustrial, pero sobre todo para aquellos sectores que disponen de un acotado acceso a la información. Adicionalmente, identificar las variables que explican parcialmente las oscilaciones de las cotizaciones, resultaría un aporte de gran magnitud dado que permitiría expandir la agenda futura de investigación en base a ellos.

5. ALCANCE Y LIMITACIONES

La identificación y cuantificación del impacto de los factores exógenos que afectan las cotizaciones del precio internacional de la Harina de Soja se analizará mediante el empleo de fuentes de información de acceso público, con el objetivo de facilitar su replicación y profundizar su posterior análisis.

A su vez, es necesario remarcar que la información generada en el presente trabajo no debería ser utilizada como única estrategia para el manejo de riesgo asociado a las fluctuaciones de precios, sino como un componente adicional dentro de la misma.

6. HIPÓTESIS

Mediante el uso de Redes Neuronales Recurrentes (LSTM específicamente) se podría predecir la fluctuación futura de los precios de la Harina de Soja en el mercado de Chicago (CBOT).

7. OBJETIVOS

El objetivo del presente trabajo consiste en desarrollar un modelo predictivo de la tendencia del precio diario de la Harina de Soja (CBOT) a través de la utilización de Redes Neuronales Recurrentes (RNR), con una precisión aceptable para determinar tendencias de corto y mediano plazo.

Como objetivos secundarios, este estudio pretende elaborar un modelo de predicción de las tendencias de precio mediante el uso de modelos auto regresivos integrados de media móvil (ARIMA) para su posterior comparación con RNR.

8. METODOLOGÍA

8.1. TECNICAS DE MACHINE LEARNING

En primera instancia se analizará la serie de precios diaria de la Harina de Soja mediante el uso de modelos auto regresivos integrados de media móvil (ARIMA).

En segundo lugar, se desarrollará un modelo predictivo a partir del uso de una Red Neuronal Recurrente (del tipo LSTM) utilizando las variables enunciadas anteriormente.

Finalmente, se compararán ambos modelos a través del Error Medio Absoluto (MAE) a fin de determinar la superioridad de uno frente a otro.

8.2. HERRAMIENTAS

El trabajo se realizará utilizando Jupyter Notebooks en lenguaje Python 3.7 [24] corriendo en un entorno virtual a través de Google Colab [15]. Las librerías básicas para correr los modelos de Machine Learning serán Keras [16]/Tensorflow [23].

8.2.1. MODELO AUTORREGRESIVO DE MEDIA MÓVIL (ARIMA)

El modelo ARIMA es una metodología econométrica basada en modelos dinámicos que utiliza datos de series temporales (Box et al., 1970) desde un punto de vista estocástico. Estos modelos consideran la dependencia existente entre los distintos datos, esto es, cada observación en un momento dado es modelada en función de los valores anteriores. ARIMA permite describir un valor como una función lineal de datos previos y errores debidos al azar pudiendo, además, incluir un componente cíclico o estacional.

Este tipo de modelos resultan adecuados para predicciones en el corto plazo (dada su capacidad de aprender de los cambios en la dinámica de la serie) al tiempo que son económicos y rápidos de construir. Por otro lado, los resultados obtenidos suelen ser pobres para predecir en momentos de cambios estructurales y puntos de giro, el horizonte de pronóstico es corto frecuentemente y son menos eficientes en el uso de la información.

8.2.2. RED NEURONAL RECURRENTE

Una Red Neuronal Artificial (RNA) es un modelo simplificado que imita el modo en que el cerebro humano procesa la información a través del funcionamiento simultáneo de un número elevado de unidades de procesamiento interconectadas similares a versiones abstractas de neuronas (IBM [19], Figura 1). Las RNA se aplican en un amplio campo de acción que abarca el análisis y procesamiento de señales, reconocimientos de imágenes, diagnósticos médicos, robótica y procesamiento del lenguaje.

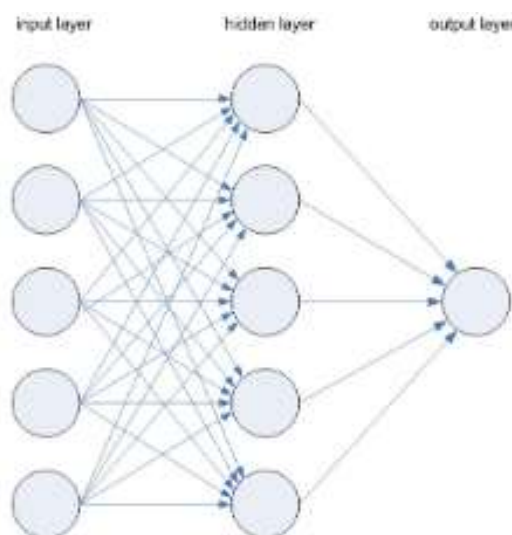


Figura 1. Estructura de una Red Neuronal Artificial (RNA, IBM, 2021).

Sin embargo, una de las principales limitaciones de las RNA radica en que no poseen memoria asociada al modelo, lo que

resulta problemático para el análisis de datos secuenciales o series de tiempo. Zhang and Qi (2005) demostraron que las redes neuronales no son capaces de capturar la estacionalidad por default, al tiempo que la des-estacionalización y la desintegración de la tendencia pueden mejorar su capacidad predictiva.

Las Redes Neuronales Recurrentes (RNR) incluyen en su diseño una mirada retrospectiva que cumple una función de memoria. Como resultado, no sólo los datos de entrada son empleados en la elaboración de la predicción, sino también los datos generados por la RNR.

A pesar de esto, las RNR no siempre se comportan dentro de lo esperado ya que su proceso de entrenamiento es complejo (Petneházi, 2019). Al mismo tiempo, las dependencias temporales con distintas características y las dependencias de largo plazo no son capturadas adecuadamente por las RNR debido al efecto del gradiente descendiente (Lindemann, et al., 2021).

8.2.3 ARQUITECTURA LSTM

Las Redes Neuronales Long Short-term Memory (LSTM, Hochreiter and Schmidhuber, 1997) consisten en un tipo de arquitectura de RNR propuestas como solución al problema del gradiente descendiente y, posteriormente, mejoradas al introducir una compuerta adicional (forget) en su estructura. Una red LSTM utiliza neuronas que permiten decidir qué información se descarta y cuál se conserva, en función del estado actual.

Éstas disponen de tres puertos o compuertas, i.e. input (permite añadir nuevos elementos de la memoria), forget (permite eliminar elementos de la memoria) y output (permite crear el estado oculto actualizado), que permiten modificar el estado de un vector que es propagado iterativamente para capturar las dependencias de largo plazo (Figura 2).

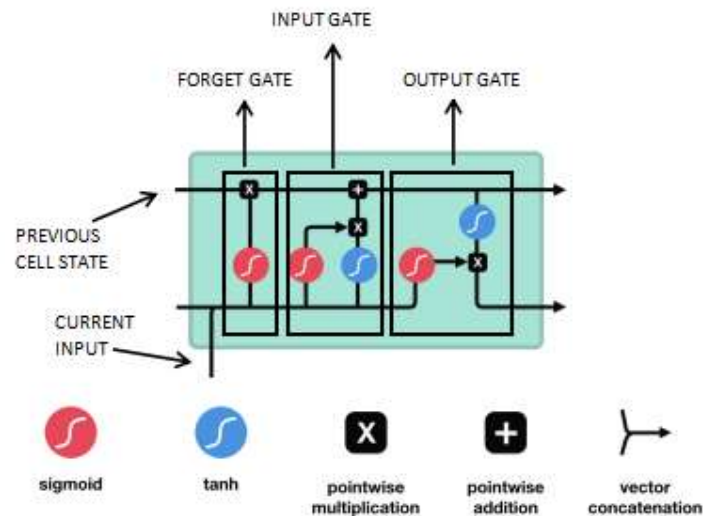


Figura 2. Estructura de una Célula LSTM (Analytics Vidhya, 2021).

9. EXPERIMENTACIÓN

9.1. DEFINICIÓN DE VARIABLES

Los distintos modelos a desarrollar serán generados a partir de la utilización de las siguientes variables:

- Serie de precios diaria de la Harina de Soja en el mercado de Chicago (CBOT), siempre considerando la posición más cercana. Precio de Apertura, Cierre, Máximo y Mínimo.
- Posición de mercado de grandes fondos (Managed Money Funds) sobre contratos de Harina, Aceite y Poroto de Soja.
- Índices de Vegetación (NDVI) generados a partir de sensores remotos (MODIS, NASA).
- Estimaciones de stocks o inventarios disponibles en principales países productores, i.e. Argentina, Brasil y Estados Unidos.
- Estimaciones de producción de Soja en principales países productores, i.e. Argentina, Brasil y Estados Unidos.

9.2. OBTENCION DE LOS DATOS

9.2.1. SERIE DE PRECIOS DE HARINA DE SOJA.

La serie histórica de precio diario de Harina de Soja se extrajo de la plataforma Yahoo Finance [31] utilizando la librería `pandas_datareader` para el período 13/Jun/2006 al 8/Abr/2022.

Se consideraron los valores diarios de Cierre (Last) y la variable se definió de la siguiente manera:

TARGET: Precio de Cierre de la Harina de Soja (Variable Objetivo).

9.2.2. POSICION DE FONDOS (MMF).

La posición de mercado de los grandes fondos (Managed Money Funds) se extrajo del reporte semanal provisto por la Commodity Futures Trading Commision (CTFC) [7], en el cual se detalla la cantidad de contratos comprados y vendidos de Poroto, Harina y Aceite de Soja por los distintos integrantes del mercado. Se considera la posición neta (contratos comprados – contratos vendidos).

La serie temporal de datos semanal se transformó a diaria interpolando los valores entre los períodos reportados.

Las variables resultantes fueron las siguientes:

MM_SOJA: Cantidad de contratos comprados – Cantidad de contratos vendidos de poroto de Soja.

MM_SBO: Cantidad de contratos comprados – Cantidad de contratos vendidos de Aceite de Soja.

MM_SBM: Cantidad de contratos comprados – Cantidad de contratos vendidos de Harina de Soja.

9.2.3. INDICES DE VEGETACION.

Los datos referidos al Índice de Vegetación (NDVI) de cultivos en Sudamérica y Estados Unidos son provistos por la plataforma GIMMS (Global Agricultural Monitoring, NASA) [13] a través del sensor Terra MODIS, con una resolución temporal de 8 días. Los datos diarios fueron generados a través de la interpolación de los valores reportados.

Los valores de NDVI considerados representan un promedio de las principales regiones productivas de Argentina, Brasil y USA.

Las variables fueron definidas según la siguiente nomenclatura,

ARG SAMPLE VALUE: Valor medio de NDVI para una fecha determinada en la región productiva de Argentina.

ARG ANOM VALUE: Anomalía (respecto de valores normales) del valor medido de NDVI en Argentina.

ARG MIN VALUE: Valor mínimo de NDVI para una fecha determinada en la región productiva de Argentina.

ARG MAX VALUE: Valor MÁXIMO de NDVI para una fecha determinada en la región productiva de Argentina.

Análogamente las mencionadas variables fueron generadas para Brasil y USA.

9.2.4. STOCKS DE HARINA, ACEITE Y POROTO DE SOJA EN USA Y PRODUCCIÓN DE SOJA EN ARGENTINA, BRAZIL Y USA.

Las estimaciones de inventarios de Poroto, Harina y Aceite de Soja de USA y los niveles proyectados de producción se tomaron de la base de datos del USDA (United States Department of Agriculture, s.f.). Debido a que este reporte es publicado con una frecuencia mensual, los valores de los días posteriores a la fecha de publicación fueron repetidos hasta la publicación de un nuevo informe (forward filling).

Las variables generadas fueron las siguientes:

US MEAL STOCKS: Existencias de Harina de Soja en USA.

US OIL STOCKS: Existencias de Aceite de Soja en USA.

US SOYBEANS STOCKS: Existencias de poroto de Soja en USA.

US AREA PLANTED: Área sembrada de Soja en USA.

US PRODUCTION: Producción de Soja en USA.

US YIELD: Rendimiento del cultivo de Soja en USA.

CROP ARGENTINA: Producción de Soja en Argentina.

CROP BRAZIL: Producción de Soja en Brasil.

CROP SAM: CROP ARGENTINA + CROP BRAZIL.

CROP BIG 3: CROP SAM + US PRODUCTION.

En el Cuadro 1 se detallan las fuentes de información utilizadas y las transformaciones necesarias para estandarizar la serie temporal diaria.

	FUENTE DE DATOS	FRECUENCIA	TRANSFORMACIÓN
PRECIOS COMMODITIES	CBOT	DIARIA	-
POSICION FONDOS (MM)	CTFC	SEMANAL	INTERPOLACIÓN
INDICE DE VEGETACION (NDVI)	NASA	8 DÍAS	INTERPOLACIÓN
STOCKS/PRODUCCION COMMODITIES	USDA	MENSUAL	FORWARD FILLING

Cuadro 1. Fuentes de datos utilizados y transformación necesaria para homogeneizar las series temporales.

9.2.5. DATASET FINAL

Todas las variables mencionadas anteriormente fueron compiladas en un único set de datos mediante la utilización de las librerías Pandas y Numpy de Python, **con un total de 26 variables y 3.989 registros.**

La variable a predecir es el precio de Cierre de la Harina de Soja en CBOT (TARGET).

9.3. ANÁLISIS EXPLORATORIO

9.3.1. Análisis de Correlación de los datos

Se realizó un análisis de Correlación mediante la librería pandas de Python (utilizando el método de Pearson) de las variables analizadas (cuyo gráfico puede verse en la Figura 3) observándose que:

- La posición neta de los grandes fondos de inversión (MM) en contratos de SBS y SBM tiene una correlación positiva con los precios de la Harina de Soja (0.55 y 0.60, respectivamente), esto es, a mayor cantidad de contratos de Soja y Harina de Soja comprados por los fondos, mayor es el precio en CBOT de ésta última.
- Los stocks reportados por el USDA de Soja y Aceite de Soja en USA se correlacionan negativamente con los precios de SBM (-0.43 en ambos casos). Esta relación resulta esperable debido a que, en el análisis fundamental de los mercados de commodities, cuanto menores son los niveles de stocks mayores son los precios reflejados en el mercado (principio de Oferta y Demanda).
- El nivel de producción de Soja en Argentina está negativamente relacionado con los precios de SBM en CBOT (-0.27). Este factor puede explicarse dado que el saldo exportable de dicha nación constituye aproximadamente el 40% del comercio global del mismo (USDA, 2021).

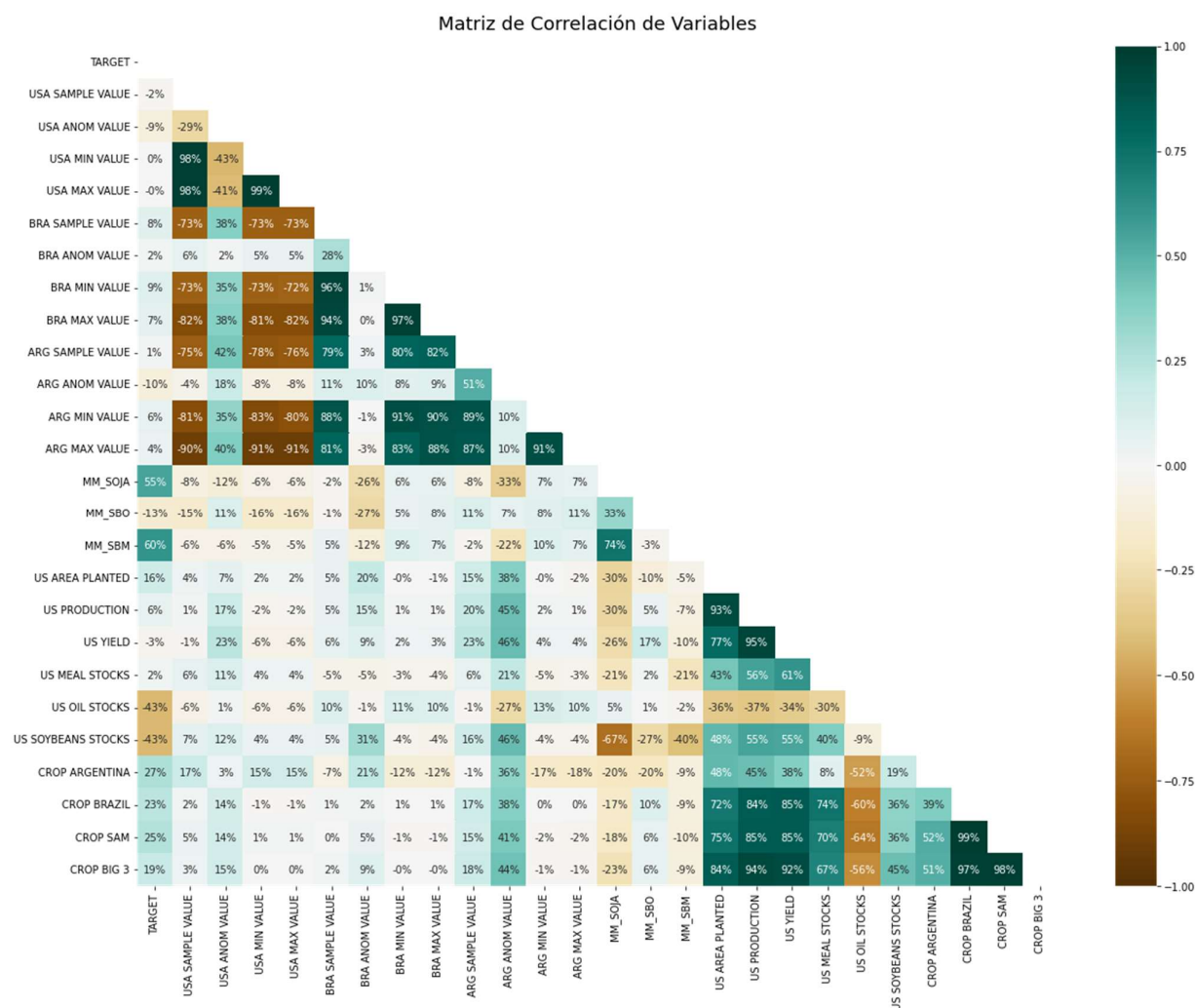


Figura 3. Matriz de Correlación de las variables analizadas.

9.3.2. Análisis de serie de tiempo (Target)

En primera instancia, se analizó la serie de tiempo de los precios de la Harina de Soja en CBOT (variable respuesta) a fin de detectar posibles estacionalidades o tendencias generales del conjunto de datos.

Como se observa en la Figura 4, no se aprecia una estacionalidad definida en la serie de precios. Este hecho resulta esperable ya que la variación en las cotizaciones no depende únicamente de factores estacionales, como los ciclos de siembra y cosecha del cultivo de Soja.

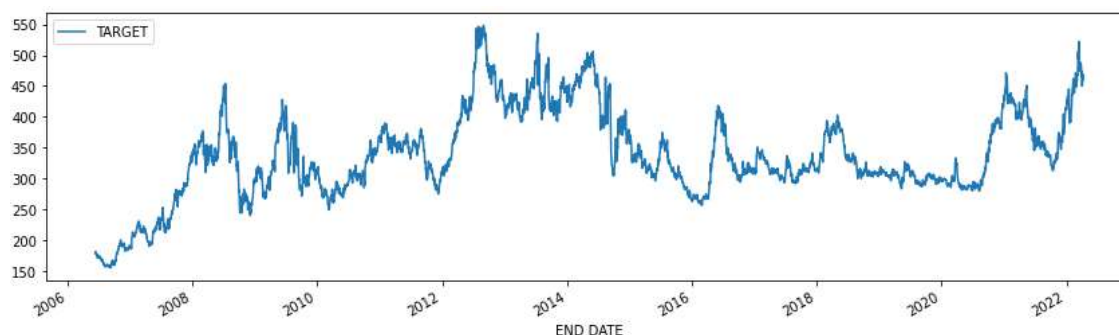


Figura 4. Serie de precios diaria del Harina de Soja en CBOT.

Al descomponer la serie (Figura 5) se observa una tendencia al alza hasta el año 2014, pero luego la misma se mantiene prácticamente constante, por lo que los datos no tienen una tendencia marcada o definida.

En términos de estacionalidad, se evidencia un claro patrón intra-anual. Una explicación potencial a este patrón podría radicar en que el mercado de Chicago, al ser una referencia global del precio de la Harina de Soja, evidencia anualmente los ciclos de producción agrícola, sobre todo de Estados Unidos con un comportamiento similar al de Oferta y Demanda. Esto es, durante la época de cosecha de Soja en el Hemisferio Norte (Septiembre/Octubre en base a estadísticas del USDA [30]), la mayor oferta de producto (Harina de Soja) tiene un impacto negativo en los precios de ese período. A medida que nos alejamos del mismo, los precios tienden a crecer por la menor disponibilidad de producto.

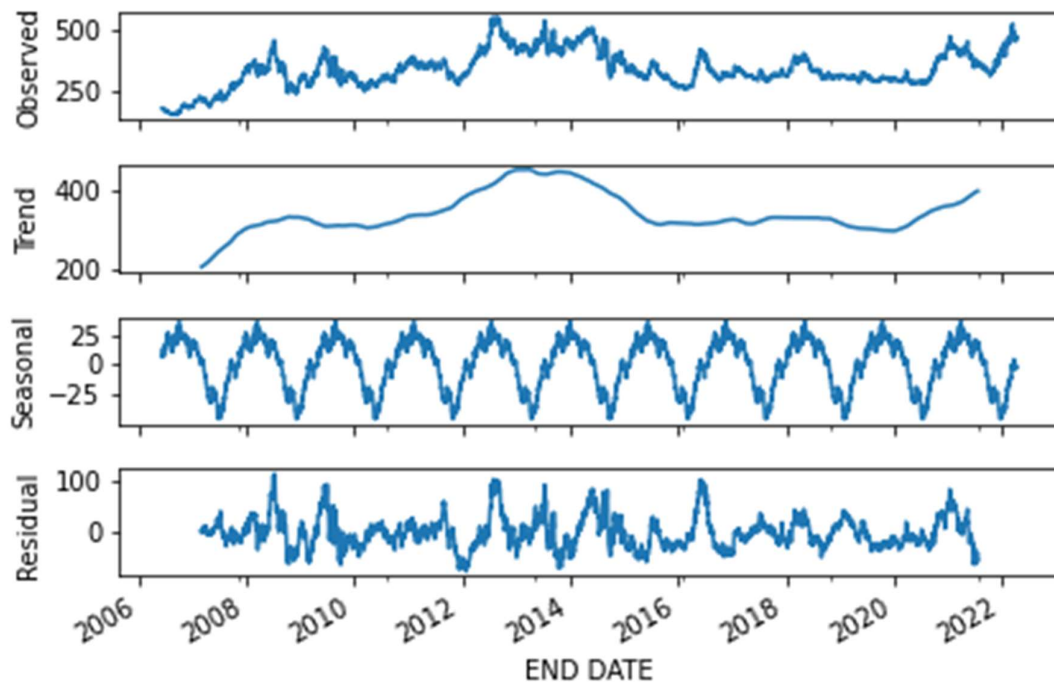


Figura 5. Serie de precios diaria del Harina de Soja en CBOT descompuesta en tendencia, estacionalidad y residuos.

A partir de la Autocorrelación parcial de la serie de datos (Figura 6) podemos afirmar que, a pesar de lo observado previamente en la descomposición de la serie, no existe una estacionalidad definida a medida que avanzamos en el tiempo (el desfase 0 siempre da 1 porque es contra sí misma sin desfase).

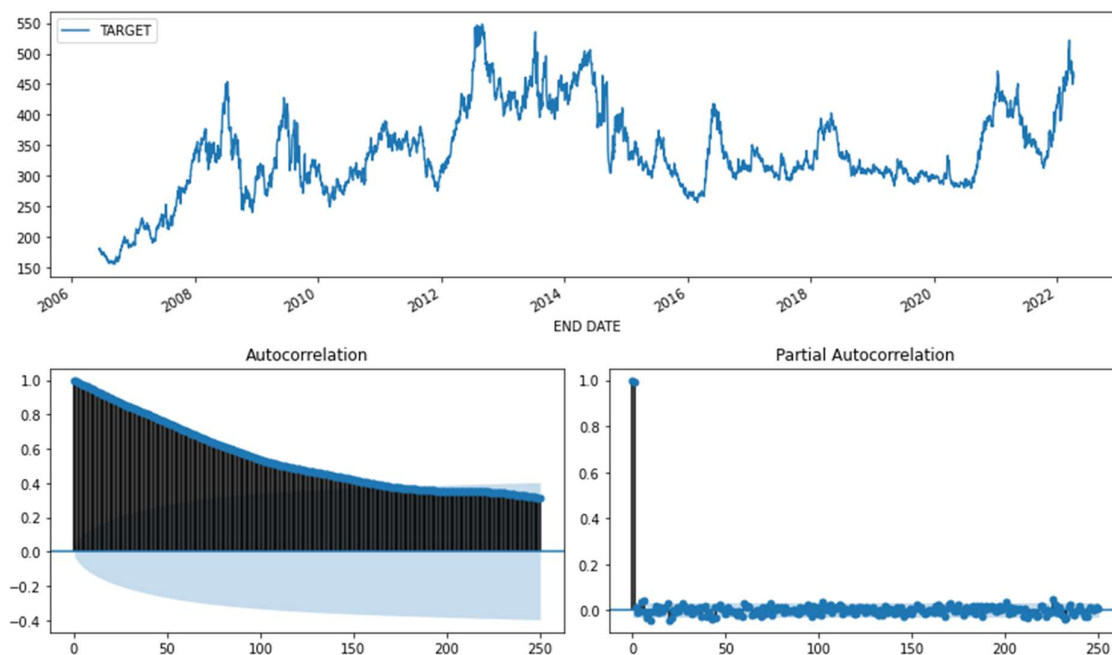


Figura 6. Autocorrelación (gráfico inferior izquierdo) y Autocorrelación Parcial (gráfico inferior derecho) de la serie de datos.

Por último, se analizó la Estacionariedad de la serie de datos, esto es, si las propiedades básicas de la distribución como la media, varianza y covarianza son constantes en el tiempo. A tal fin se empleó la prueba de Dickey-Fuller, que consiste en una prueba de raíz única que detecta estadísticamente la presencia de conducta tendencial estocástica en las series temporales de las variables mediante un contraste de hipótesis:

Hipótesis Nula (H_0): La serie de datos es No estacionaria.

Hipótesis Alternativa (H_1): La serie de datos es Estacionaria.

El valor p obtenido a través de esta prueba resultó **0.0213 (< 5%)** por lo que se rechazó la H_0 concluyéndose que la serie de datos bajo análisis es estacionaria.

9.4. CONJUNTO DE ENTRENAMIENTO Y TESTEO

Para el entrenamiento de los distintos modelos de análisis de series temporales se dividió el conjunto de datos en dos sets, i.e. entrenamiento y testeo. El primero contiene el 80% de los datos, mientras que el segundo dispone del 20% remanente (Figura 7).

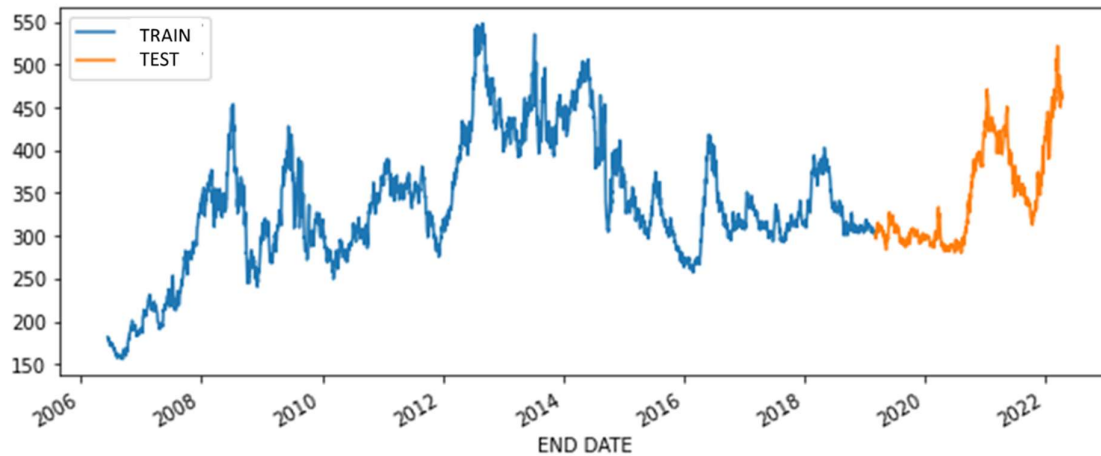


Figura 7. Conjunto de Entrenamiento (80% de los datos) y Testeo (20%).

9.5. MODELO ARIMA

Para la elección del modelo ARIMA se utilizó el módulo Auto.ARIMA de la librería SKTIME, el cual selecciona el mejor modelo a través de la optimización del valor AIC (Akaike Information Criterion), esto es, minimizándolo.

En la Figura 8 se puede observar el resultado de la predicción para el conjunto de Testeo a partir del mejor modelo (mínimo AIC), cuyo MAE resultó de **44.33**.

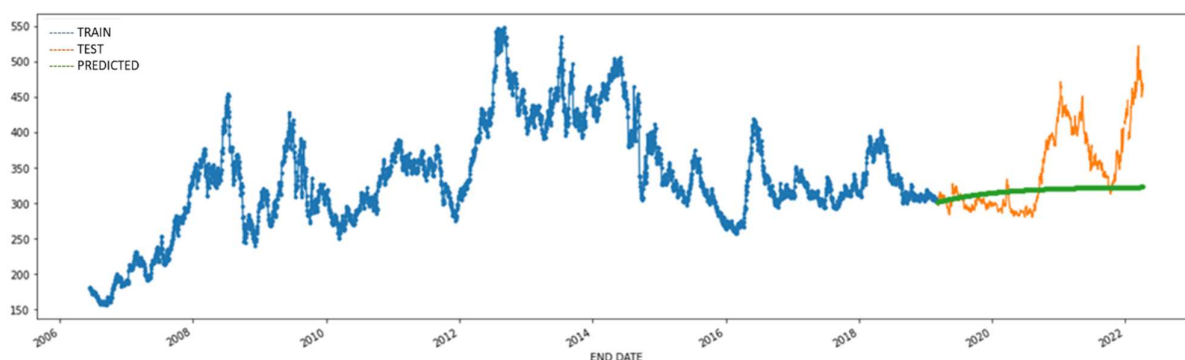


Figura 8. Predicción del modelo ARIMA para el conjunto de datos de Testeo.

A continuación, se generó modelo a partir de RNR (LSTM) con el objeto de comparar su performance relativa respecto del modelo ARIMA descrito anteriormente.

9.6. MODELO LSTM

La implementación de esta Red Neuronal se efectuó a través Google Colab mediante de la utilización de la librería Tensorflow con Keras como interface en Python. Los hiperparámetros utilizados fueron los siguientes:

- **Timesteps:** Cantidad de días previos que LSTM considera para cada punto. Se utilizaron 14 días.
- **Capas:** Se emplearon 2 capas
- **Número de épocas (epochs):** 20
- **Batch Size:** 16

A diferencia del modelo ARIMA en el cual se analizó únicamente la variable TARGET (serie de precios diarios), en este caso todas las variables adicionales a la serie de precios de la Harina de Soja fueron consideradas. Para esto, previo al entrenamiento de la Red LSTM todos los datos fueron estandarizados y escalados utilizando el módulo StandardScaler de la librería Sklearn (Python).

Las curvas de aprendizaje de los modelos durante la etapa de entrenamiento pueden utilizarse para diagnosticar posibles problemas de overfitting/underfitting. En este caso, la gráfica de pérdida de entrenamiento (Figura 9) muestra una disminución paulatina de las curvas de Training y Validation Loss hasta alcanzar un punto de relativa estabilidad, indicando un ajuste aceptable.

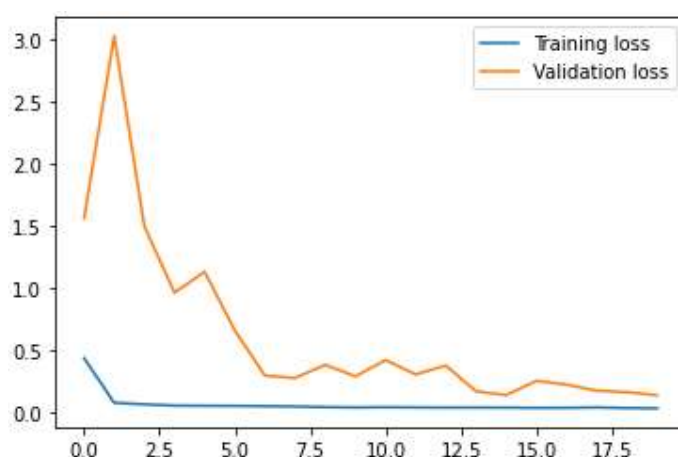


Figura 9. Curvas de Loss para el conjunto Training (azul) y validación (naranja) durante las epochs del entrenamiento del modelo LSTM.

Finalmente, con el objetivo de comparar la performance de este modelo con ARIMA, el valor MAE calculado fue de **31.04**, esto es, menor que el arrojado por el modelo previo (44.33).

Alternativamente al MAE, gráficamente (Figura 10) se puede apreciar que la Red Neuronal LSTM si bien aún no es capaz de predecir con exactitud incrementos de precios en cortos

períodos de tiempo, en general resulta una herramienta válida a fin de estimar la tendencia general del mercado.



Figura 10. Predicción de los precios de Harina de Soja en CBOT en base al modelo LSTM (línea roja) y valores observados reales (línea negra).

9.7. MEDIA MÓVIL 20 DÍAS

Una forma alternativa de analizar los resultados consiste en comparar la media móvil de 20 días en lugar de los valores diarios. Este es un indicador técnico muy utilizado por inversores y traders al momento de determinar la dirección de la tendencia de corto plazo de distintos subyacentes.

De esta manera, se pueden observar con mayor nitidez las tendencias a corto plazo disminuyendo las distorsiones generadas por precios diarios con extrema volatilidad resultantes, por ejemplo, de eventos inesperados.

En la Figura 12 se puede apreciar como la curva de precios pronosticados por el modelo reconoce las tendencias alcistas y bajistas del mercado durante gran parte del período analizado.

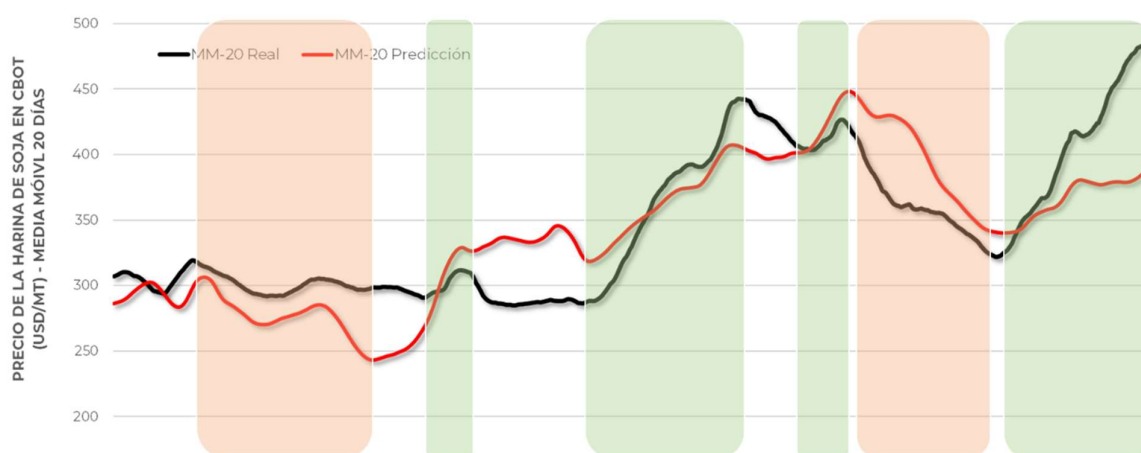


Figura 11. Media Móvil de 20 días de la Predicción de los precios de Harina de Soja en CBOT en base al modelo LSTM (línea roja) y valores observados reales (línea negra). Sombreado verde destaca períodos de coincidencia en la tendencia al alza de las predicciones y los valores reales, mientras que sombreado rojo coincidencia en períodos bajistas.

10. RESULTADOS

A partir del análisis exploratorio de los datos se observó que los precios de la Harina de Soja en CBOT están positivamente correlacionados con la posición neta de los grandes fondos de inversión (MM) en contratos de poroto de Soja y Harina de Soja (0.55 y 0.60 respectivamente). Esto podría explicarse por la gran magnitud y escala de las operaciones de estos actores, que producen fluctuaciones significativas en las cotizaciones diarias.

En menor medida, los stocks de Soja (-0.43) y el nivel de producción de este grano en Argentina (-0.27) también mostraron correlaciones negativas con la variable respuesta. En ambos casos, la menor oferta (o disponibilidad) de poroto de Soja genera un incremento de las cotizaciones, respondiendo al principio económico de oferta y demanda.

Se realizaron dos modelos, ARIMA y RNR LSTM, con el objeto de comparar la performance relativa de ambos. En base al Error Medio Absoluto (MAE por sus siglas en inglés) el modelo de RNR LSTM resultó superior dado su MAE de **31.04**, mientras que ARIMA arrojó un valor de **44.33**.

Adicionalmente a esta métrica de error, gráficamente los resultados obtenidos por la RNR LSTM resultan superiores permitiendo identificar las tendencias a corto y mediano plazo de los precios de la Harina de Soja. Esto último, puede corroborarse a través de la transformación de la serie de precios diaria a una media móvil de 20 días. De esta manera, las variaciones diarias de precios son atenuadas, permitiendo reconocer las distintas tendencias con mayor grado de nitidez. Como presentó anteriormente, la RNR LSTM demuestra una adecuada capacidad predictiva de las tendencias alcistas y bajistas de las cotizaciones de la Harina de Soja.

Analizando los resultados obtenidos es posible corroborar que, bajo los parámetros definidos en el presente trabajo, las RNR demuestran mejores resultados al momento de predecir tendencias futuras del precio de la Harina de Soja.

11. CONCLUSIONES

Existen numerosos modelos utilizados para el análisis de distintas series temporales. Sin embargo, al estar influenciadas por numerosos factores, las series de precios de commodities agrícolas constituyen un desafío particular. En el presente trabajo se comparó la performance predictiva relativa de un modelo ARIMA basado únicamente en la serie temporal de precios y una Red Neuronal Recurrente del tipo LSTM con variables exógenas, sobre los precios de la Harina de Soja en CBOT.

A partir de los resultados obtenidos, podemos inferir que modelos de Redes Neuronales Recurrentes del tipo LSTM empleando variables exógenas a la serie de precios resultan un enfoque superador en comparación con métodos Autoregresivos de Media Móvil basados únicamente en la serie de datos, a partir de las métricas de error calculadas (MAE 31.04 en RNR LSTM vs 44.33 en ARIMA).

A la vez, gráficamente las RNR LSM también demuestran una mayor capacidad predictiva, sobre todo mediante la transformación de la serie de precios diaria a una media móvil de 20 días. Como consecuencia, las variaciones de precios diarios son atenuados, permitiendo una mejor identificación de las tendencias alcistas o bajistas de las cotizaciones. Durante el período analizado, la RNR LSTM generada durante el presente trabajo reconoció exitosamente las tendencias observadas.

Como se ha comprobado, las RNR del tipo LSTM han demostrado una aceptable precisión en su nivel predictivo. Sin embargo, a los fines de predecir futuras tendencias de precios a partir de este modelo, se deben generar predicciones de cada una de las variables exógenas empleadas en el mismo, como por ejemplo la evolución de la posición de los fondos de inversión o el Índice de Vegetación (NDVI) de Argentina, Brasil y EEUU. Este factor impone, en cierta medida, restricciones al momento de utilizar esta herramienta y subraya la necesidad de trabajar con distintos escenarios futuros para cada una de las variables utilizadas.

11.1. TRABAJOS FUTUROS

En el marco de futuros trabajos que amplíen el espectro abarcado en el presente estudio, podrían enmarcarse:

- El primero sugiere explorar variables exógenas adicionales a incluir en el análisis, como por ejemplo Índices de Precios de commodities energéticos. Por ejemplo, Chiou-Wei et al (2019) demostraron la existencia de distintos grados de correlación entre los precios del Petróleo, Gas Natural, Maíz, Soja y Etanol.
- En segundo lugar, la generación de modelos externos que permitan predecir el comportamiento de las variables empleadas en este modelo (como, por ejemplo, el Índice de Vegetación) y permitan extender el período predictivo podría resultar de interés. En el presente trabajo, la RNR LSTM entrena con los últimos 14 días prediciendo el próximo valor. La disponibilidad de modelos que permitan estimar el comportamiento de estas variables a futuro (con alta precisión) permitirían extender el período de predicción más allá de únicamente un día.

12. REFERENCIAS BIBLIOGRÁFICAS

1. Analytics Vidhya, 2021. <https://medium.com/analytics-vidhya>.
2. Arrieta, José; Lescano, D.; Mirada, M. ; Vitale, L. ; Felicioni, E. ; Vaudagna, J. 2015. Estrategias nutricionales y de alimentación en recría y engorde. Impacto en los costos. Fericerdo 2015.
3. Benjamin Lindemann, Timo Müller, Hannes Vietz, Nasser Jazdi, Michael Weyrich. A Survey on long short-term memory networks for time series prediction. 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, CIRP ICME '20. Institute of Industrial Automation and Software Engineering, University of Stuttgart, Pfaffenwaldring 47, 70569 Stuttgart, Germany. 2021.
4. Bessler, D. A. Adaptive expectations, the exponentially weighted forecast and optimal statistical predictors: A revisit. Agricultural Economics Research, 34, 16–23. 1982.
5. Borja y Medel. Avances en la alimentación del porcino. Avances en nutrición y alimentación animal. 1(1):63. 1998.
6. Box, G. & G. M. Jenkins. Time-series Analysis, Forecasting and control. San Francisco, California: Holden-Day. 1970.
7. Commodity Futures Trading Commission. www.cftc.gov.
8. CME Group. Commodities Mercantile Exchange. <https://www.cmegroup.com>.
9. Chiou-Wie, S., Chen S. & Zhu Z. Energy and Agricultural Commodity Markets interaction: An analysis of Crude Oil, Natural Gas, Corn, Soybeans, and Ethanol prices. The Energy Journal, Vol. 40, No. 2. 2019.

10. Dirección de Mercados Agropecuarios.
www.magyp.gob.ar/sitio/areas/ss_mercados_agropecuarios.
11. G Peter Zhang and Min Qi. Neural network forecasting for seasonal and trend time series. *European journal of operational research*, 160(2):501–514, 2005.
12. Gábor Petneházi. Recurrent Neural Networks for Time Series Forecasting. Doctoral School of Mathematical and Computational Sciences University of Debrecen. University of Debrecen. 2019.
13. GIMMS. Global Agricultural Monitoring. NASA.
www.glam1.gsfc.nasa.gov.
14. Gómez, D., Salvador, P., Sanz, J. & Casanova, J. L. Modelling wheat yield with antecedent information, satellite and climate data using machine learning methods in Mexico. *Agricultural and forest meteorology*, 300, 108317. 2021.
15. Google Colaboratory. <https://colab.research.google.com/>
16. Gulli, A., & Pal, S. *Deep learning with Keras*. Packt Publishing Ltd. 2017.
17. H.C. de Boer, M.M. van Krimpen, H. Blonk, M. Tyszler. Replacement of soybean meal in compound feed by European protein sources. *Livestock Research Report* 819. Wageningen UR Livestock Research. 2014.
18. Hochreiter, S. & Schmidhuber, J. Long Short-term Memory. *Neural Computation*, 9 (8). Pages 1735-1780. 1997.
19. IBM, “El Modelo de redes neuronales”.
<https://www.ibm.com/docs/es/spss-modeler/saas?topic=network-neural-model>. 2021.

20. Khamis, A., & Abdullah, S. Forecasting wheat price using backpropagation and narx neural network. The international journal of Engineering and Science, 3(11), 19-26. 2014.
21. Lindemann, B., Müller, T., Vietz, H., Jazdi, N. & Weyrich, M. A survey on long short-term memory networks for time series prediction. Proedia CIRP. Volume 99. Pages 650-655. 2021.
22. Lopes, L. P. Prediction of the Brazilian natural coffee price through statistical machine learning models. SIGMAE, 7(1), 1-16. 2018.
23. Martín Abadi, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. 2015.
24. Montaña Moreno, Juan José. Redes Neuronales Artificiales aplicadas al análisis de datos. Tesis Doctoral. Universidad de Les Illes Balears. 2002.
25. Petneházi, G. Recurrent Neural Networks for Time Series Forecasting. Doctoral School of Mathematical and computational Sciences. University of Debrecen. 2019.
26. Python Software Foundation. (2021). Python. <https://www.python.org/>
27. Racine Ly, Fousseini Traore and Khadim Dia. Forecasting commodity prices using long-short memory neural networks. 2021.
28. Ribeiro, M. H. D. M., & Dos Santos Coelho, L. Ensemble approach based on bagging, boosting, and stacking for short-term prediction in agribusiness time series. Applied soft computing, 86, 105837. 2020.

29. Timmermann, A. Forecast combinations. Handbook of Economic Forecasting, 1, 135–196. 2006.
30. USDA. Foreign Agricultural Service. <https://apps.fas.usda.gov/psdonline>. 2021.
31. Wang, Z., & Bessler, D. A.. Forecasting performance of multivariate time series models with full and reduced rank: An empirical examination. International Journal of Forecasting, 20(4), 683–695. 2004.
32. Xiaojie Xu, Yun Zhang. Soybean and Soybean Oil Price Forecasting through the Non-linear Autoregressive neural network (NARNN) and NARNN with Exogenous inputs (NARNN-X). Intelligent Systems with Applications. 2022.
33. Yahoo Finance. www.finance.yahoo.com.
34. Yang, J., & Awokuse, T. O. Asset storability and hedging effectiveness in commodity futures markets. Applied Economic letters, 10 (8), 487-491. 2003.
35. Zhang P & Qi, M. Neural network forecasting for seasonal and trend time series. European Journal of Operational Research. Volume 160, Issue 2. Pages 501-514. 2005.