

Tovar, Daniel Federico. *Estudiante Maestría en Inteligencia Artificial*. Gómez Parra, Héctor., *Estudiante Maestría en Inteligencia Artificial*

# Generación de un modelo para predecir el precio del café (KC=F) en la bolsa de Nueva York

**Abstract—** This study explores various machine learning approaches and feature engineering techniques to optimize the prediction of stock prices in the market. Multiple models, including neural networks, decision trees, and others, would be used to compare based on their predictive performance. Hyperparameter tuning is going to also be analyzed, and the models' decisions are going to be justified using performance metrics and loss visualizations. Additionally, we are going to use features about temperature, exports (offer and demand) to have more specific but reliable models. The research aims to identify the most effective model for accurately forecasting coffee prices in the dynamic stock market.

**Resumen—** Este estudio explora varios enfoques de aprendizaje automático y técnicas de ingeniería de características para optimizar la predicción de los precios de las acciones en el mercado. Se utilizarán múltiples modelos, incluidas redes neuronales, árboles de decisión y otros, para compararlos en función de su rendimiento predictivo. También se analizará el ajuste de hiperparámetros y se justificarán las decisiones de los modelos mediante métricas de rendimiento y visualizaciones de pérdidas. Además, vamos a utilizar características sobre temperatura, exportaciones (oferta y demanda) para tener modelos más específicos pero fiables. El objetivo de la investigación es identificar el modelo más eficaz para predecir con exactitud los precios del café en el mercado dinámico de valores.

*Index Terms:* Stock market forecasting, Market volatility, Weather patterns, Coffee price prediction, Environmental impact

## I. PROBLEMÁTICA

El cambio climático y su impacto en la industria del café han sido objeto de una creciente atención. Las alteraciones en los patrones climáticos, como el aumento de las temperaturas, la variabilidad de las lluvias y eventos climáticos extremos, están afectando significativamente la producción y la calidad del café en todo el mundo. Estudios recientes han señalado que el cambio climático está provocando una serie de problemas en la cadena de suministro del café.

Un artículo de la revista *Frontiers in Plant Science* (Ahmed et al., 2021) destacó que las elevadas temperaturas están llevando a un aumento de las enfermedades del café, como la roya, que puede diezmar las cosechas. Además, una investigación publicada en la revista *Nature Plants* (Gómez-Dans et al., 2019) subrayó la influencia de las sequías y las variaciones en las precipitaciones en las regiones cafetaleras de América Latina, lo que conlleva una disminución en la producción de café.

La volatilidad de los precios del café en los mercados internacionales también ha sido ampliamente discutida. Un estudio de mercado publicado por la Organización Internacional del Café (ICO, 2021) señaló que "la variabilidad climática puede dar lugar a fluctuaciones en la producción, lo que a su vez afecta la oferta y la demanda en los mercados internacionales". En respuesta a estos desafíos, el presente proyecto se centra en analizar datos históricos de precios de café y condiciones climáticas en la región de Brasil, la principal productora de café del mundo. Nuestra iniciativa busca desarrollar un modelo de predicción que permita comprender en mayor profundidad la relación entre el cambio climático y las fluctuaciones en los precios del café en la Bolsa de Nueva York.

## II. INTRODUCCIÓN

El mercado global del café, un producto de enorme significado económico y cultural está intrínsecamente ligado a variabilidades climáticas, especialmente en regiones productoras clave como Brasil. Este estudio se enfoca en la predicción del comportamiento del precio internacional del café, considerando variables climáticas como heladas y precipitaciones en las áreas productoras de café de Brasil. El período de análisis se extiende desde finales de 2010 hasta finales de 2022, una ventana temporal que abarca eventos climáticos significativos y fluctuaciones de mercado relevantes.

La predicción precisa de los precios del café tiene implicaciones sustanciales tanto para la economía de los países productores como para la cadena global de suministro del café. En este contexto, aplicamos tres modelos de aprendizaje de máquina: Redes Neuronales Recurrentes, Extreme Gradient Boosting y Regresión Lineal. Cada uno de estos modelos ofrece enfoques únicos y complementarios para abordar la complejidad de predecir los precios del café, un problema caracterizado por su alta volatilidad y la influencia de numerosos factores exógenos.

Las Redes Neuronales son adecuadas para capturar patrones complejos y no lineales en grandes conjuntos de datos, lo que las hace particularmente útiles para analizar las relaciones entre las condiciones climáticas y los precios del café. Por otro lado, el modelo Extreme Gradient Boosting (XGBoost) es un algoritmo que es particularmente eficiente para manejar datos no lineales y complejos, lo que lo hace ideal para capturar las complejas relaciones entre las condiciones climáticas y los precios del café. Finalmente, la Regresión Lineal es un método estadístico que, a pesar de su simplicidad, es utilizado para comprender las relaciones entre variables. En el contexto del proyecto, puede servir como una buena línea de base para comparar la eficacia de modelos más complejos como el Extreme Gradient Boosting y las redes neuronales recurrentes, ofreciendo una visión clara de la relación lineal entre las características y el objetivo de predicción.

Este estudio no solo busca predecir si los precios del café subirán, bajarán o se mantendrán estables, sino también entender cómo las variaciones climáticas específicas en Brasil, como heladas y precipitaciones, afectan estos precios. Al hacerlo, aspiramos a contribuir con herramientas analíticas valiosas para los actores del mercado del café y a ofrecer perspectivas para una planificación más efectiva en el sector agrícola frente a los desafíos del cambio climático.

### III. METODOLOGÍA CRISP-DM

#### A. COMPRENSIÓN DEL NEGOCIO

Este estudio aborda una necesidad crítica en el mercado del café: prever los precios internacionales en el contexto de variables climáticas fluctuantes. Nuestro objetivo es proporcionar herramientas predictivas que puedan ayudar a los productores, distribuidores, inversionistas y otros actores clave del mercado a tomar decisiones informadas, mitigar riesgos y capitalizar oportunidades. Entendemos que el mercado del café es altamente sensible a factores climáticos, y nuestro enfoque se centra en desentrañar estas complejidades para predecir tendencias de precios.

#### B. COMPRENSIÓN DE LOS DATOS

Nuestro proyecto se basará en tres pilares de datos esenciales. Primero, analizaremos los precios diarios del café arábica, accediendo a los datos disponibles en Yahoo Finance y en el sitio web de la Organización Internacional del Café, específicamente en su sección "Historical Data on the Global Coffee Trade, 2023". Como segunda fuente, nos centraremos en la exportación de sacos de 60 kg de café verde, la unidad estándar para el precio del café arábica, poniendo

especial atención en los principales países productores: Colombia, Brasil y Vietnam. El tercer conjunto de datos provendrá de las temperaturas registradas en dos regiones cruciales de Brasil, Minas Gerais y Sao Paulo, que juntas contribuyen con más del 60% de la producción de café verde para exportación, según Costa (2020). Estos datos climáticos serán recabados del Instituto Nacional De Meteorología (INMET) de Brasil, que dispone de registros históricos detallados (8760 registros anuales, lo que se traduce en 24 registros diarios por región, abarcando 17 variables como temperatura y precipitaciones).

Para capturar de manera más efectiva la influencia del clima en los precios del café, planeamos superponer los datos climáticos y de precios. Esto nos permitirá mejorar la precisión de nuestros modelos y establecer correlaciones más robustas. La recolección de datos se iniciará desde finales de 2010, un año marcado por tres picos significativos en los precios del café, con especial énfasis en el pico de 2021, atribuido en parte a una severa helada en Brasil.

#### C. PREPARACIÓN DE LOS DATOS

Tenemos dos fuentes principales de datos. La primera proviene del precio del café arábico (símbolo KC=F), la cual puede descargarse directamente de Yahoo Finance. Este conjunto de datos consta de 8 columnas, en su mayoría numéricas, que son: Date, Open, High, Low, Close, Adj Close, y Volume. La columna "Date" actúa como índice, mientras que las siguientes representan el precio de apertura, el precio máximo, el precio mínimo, el cierre, el cierre ajustado, y el volumen de contratos negociados en esa fecha, respectivamente. Esta base de datos ha sido sometida a una limpieza básica, eliminando 4 valores NaN desde el año 2011.

La segunda fuente representó un desafío mayor, siendo descargada directamente del sitio web del Servicio Meteorológico Nacional de Brasil. La información estaba consolidada por ciudad, pueblo o región, y abarca desde el año 2000 hasta el 31 de octubre de 2023. Para identificar las zonas cafeteras de Brasil que producían café arábica, consultamos diversas páginas y, a partir de esa información, accedimos a Maps para determinar qué ciudades o pueblos cercanos tenían datos de temperatura. Finalmente, seleccionamos las siguientes ciudades para nuestro análisis: Varginha, Catalao, Manhuacu, Montes Claros, Patos de Minas, Franca, Casa Branca, Itapira, Bauru, Avare, Lencois, Vitoria da Conquista, Itirucu, Barreiras, Afonso Claudio, Maringá, Paraná, y Friburgo.

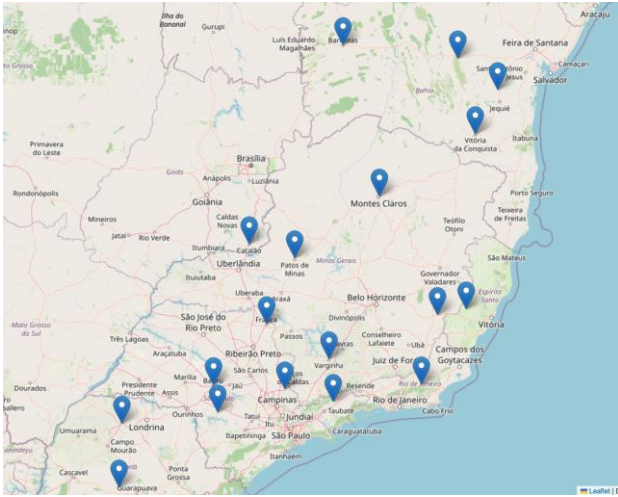


Gráfico 1: Localización regiones.

Cada ciudad o región proporcionaba un archivo CSV por año que incluía 20 columnas, y cada fila representaba una hora del día. Las columnas eran: time (UTC), total precipitation hourly (mm), atmospheric pressure at station level hourly (mb), maximum atmospheric pressure in the previous hour (aut) (mb), minimum atmospheric pressure in the previous hour (aut) (mb), global radiation (kj/m<sup>2</sup>), air temperature - dry bulb hourly (°C), dew point temperature (°C), maximum temperature in the previous hour (aut) (°C), minimum temperature in the previous hour (aut) (°C), dew point temperature maximum in the previous hour (aut) (°C), dew point temperature minimum in the previous hour (aut) (°C), maximum relative humidity in the previous hour (aut) (%), minimum relative humidity in the previous hour (aut) (%), relative humidity hourly (%), wind hourly direction (degrees), wind maximum gust (m/s), wind hourly speed (m/s), and an unnamed 19.

Limpieza de Datos: Estas variables se tradujeron del portugués al inglés para mejorar su comprensión en el análisis. Además, se revisó el porcentaje de valores perdidos, encontrándose un promedio del 9%, excepto para la variable "global radiation (kj/m<sup>2</sup>)", que presentaba un 49% de valores NaN. Por lo tanto, esta característica, junto con la unnamed 19, se eliminaron debido a la falta de información o a la presencia de muchos valores NaN. Además, para el 9% restante de las demás variables, se optó por eliminar los registros. Esto se debió a que calcular un valor promedio para su reemplazo resultaba en un gasto computacional elevado y no justificaba la inversión, ya que, en última instancia, se realizaría un promedio de promedios.

Una vez que teníamos la información cargada en dos dataframes, procedimos a realizar una concatenación para crear una única base de datos que alimentara a los modelos. La idea general era analizar cuatro valores en la información de temperatura: Temperatura bulbo seco, Mínima temperatura de la hora anterior, Presión atmosférica y Precipitación total ('air temperature - dry bulb, hourly (°C)', 'minimum temperature in the previous hour (aut) (°C)', 'total precipitation, hourly (mm)', 'atmospheric pressure at station level, hourly (mb)'). Estas cuatro columnas se añadirían a la base de datos de los precios

del café, que estaba registrada en días. Por lo tanto, fue necesario realizar un promedio para cada columna, calculando el valor promedio por hora para cada ciudad y luego promediando las horas de la madrugada (de 0.00 am a 7.00 am), ya que estas horas registran las temperaturas más bajas y, según el conocimiento empírico en Colombia, son las que más afectan a los cultivos, dado que durante este período, los cultivos no reciben luz solar para calentarse.

Con los valores promedio por ciudades y horas, los añadimos a la tabla general de precios del café. Además, llevamos a cabo una normalización de los valores, ya que teníamos características con diferentes unidades, como milímetros de lluvia, dólares de precio y temperatura en grados Celsius. Adicionalmente, eliminamos las variables de Open, Low, High y Adj Close debido a su redundancia (alta correlación), lo cual podría afectar al modelo. Así, dejamos solo la variable "Close" como etiqueta y las otras cuatro columnas de variables meteorológicas.

Después de normalizar los valores, creamos cuatro dataframes. El primero, llamado merge\_df, era el original, y luego creamos otros tres dataframes con retraso: el primero con un retraso de 1 mes, el segundo con un retraso de 6 meses y el tercero con un retraso de 12 meses. Esto se debió a nuestra consideración de que los efectos de la temperatura no se verían inmediatamente en las cosechas del mes, sino en las siguientes cosechas (teniendo en cuenta que hay dos cosechas para el café arábica por año, la normal y la "traviesa" que ocurre a los 6 meses de la cosecha normal).

#### D. METODOLOGÍA

En esta fase, aplicaremos tres modelos de aprendizaje de máquina:

- **Redes Neuronales:** Desarrollaremos una red neuronal profunda, seleccionando una arquitectura que pueda capturar dependencias temporales largas, crucial para series temporales. Experimentaremos con diferentes funciones de activación como ReLu, LeakyRelu o TanH para mejorar la captura de la dinámica no lineal en los datos y con los dataframes con diferentes lags para determinar cuál era el mejor.
- **Árboles de Decisión:** Implementaremos el modelo XGBoost, aprovechando su capacidad para manejar datos de alta dimensión y su robustez frente a overfitting. Realizaremos un ajuste fino de hiperparámetros como la tasa de aprendizaje, la profundidad máxima del árbol y el número de árboles, para optimizar el rendimiento del modelo.
- **Regresión Lineal:** Implementaremos el modelo de Regresión lineal con regularización Lasso, Ridge y Elastic Net, las cuales deberían modelos de series de tiempo más precisos que otros. Se ajustarán los valores de tasa de regularización para que el modelo reduzca el overfitting y pueda generar predicciones más estables.

## E. EXPERIMENTACIÓN

Para evaluar los modelos, utilizaremos el valor de MSE debido a que es el idóneo para comparar modelos de regresión, adicionalmente compararemos las gráficas de error vs épocas para los datos de entrenamiento y test, y la gráfica de similitud de los pronósticos vs la data real.

- **Redes Neuronales:** Tomando los aprendizajes de un modelo realizado con anterioridad cuya tarea era predecir un resultado binario de una base de datos misteriosa el cual tuvo un excelente resultado en f1-score, implementamos el mismo modelo pero ajustado a un modelo de regresión de series de tiempo, en general el modelo cuenta con un capa de entrada, cinco capas ocultas y una capa de salida, cuenta con un total de 1009 neuronas, las activaciones son por LeakyReLU y la última capa tiene una activación lineal, adicional se utiliza el optimizador Adam, un learning rate de 0.0001 y la pérdida se mide con MSE; el modelo cuenta con 50 épocas y un tamaño de batch de 34, cuenta con unos “Llamados tempranos” los cuáles rara vez se ejecutaron.

Por otro lado, se realizaron experimentos con diferentes retrasos en la data meteorológica, esto debido a que el clima de un día en específico no va a afectar el precio del café pues este ya está producido y en procesos de transformación (fermentación, secado, lavado, etc.), se utilizaron retrasos de 0, 1, 6 y 12 meses (principalmente por que el café tiene un tiempo de producción bianual generando una carga considerable una vez al año y una carga baja a mediados de año) y se comparó su desempeño para realizar otros experimentos, como por ejemplo cambiar las funciones de activación (TanH o ReLu) e inclusive ajustar la tasa de aprendizaje del modelo original.

- **Árboles de Decisión:** Teniendo en cuenta los resultados sobre retrasos del modelo de redes neuronales, y utilizando un XGBoost el cuál ha demostrado ser un excelente predictor para series de tiempo y precios con tendencia, de este modelo se se realizaron 3 corridas ajustando los valores sus parámetros, y midiendo nuevamente su desempeño a través de MSE, el modelo en general contaba con 1000 arboles una tasa de aprendizaje de 0.01 y una profundidad máxima de 12.
- **Regresión Lineal:** debido a que las series de tiempo presentan comportamientos lineales, utilizamos estos modelos regularizados para comparar el performance entre árboles y redes neuronales, se utilizaron valores de regularización de 1 y 0.5 y adicionalmente modelos tipo Ridge, Lasso y Elastic-Net, los demás parámetros se mantuvieron fijos.

## IV. RESULTADOS

- **Redes Neuronales:** Estos son los resultados del MSE con 2 corridas y para cada base de datos con diferente tipo de retraso temporal en la data meteorológica, el mejor modelo resultó con una arquitectura de 7 layers y un lag de 6 meses, esto debido a que probablemente los efectos del clima no se ven en la producción del día si no sobre los stocks para dentro de 6 meses (segunda cosecha).

MSE	RN SIN LAG	LAG 1	LAG 6	LAG 12
Corrida 1	2696.741	2577.744	1827.111	2604.480
Corrida 2	2613.772	2488.019	2210.153	2175.007
Promedio	2655.256	2532.881	2018.632	2389.744

Tabla 1:MSE por modelo RNN, Elaboración propia

Por otra parte, esta es la información de desempeño probando otras arquitecturas y tasas de aprendizaje con un Lag de 6 meses:

	LR 0.01	TanH	ReLu
MSE	3363.753	2627.712	2120.161

Tabla 2:MSE RNN Diferente Arq. Elaboración propia

Como se observa, el mejor modelo tiene una activación ReLu (Simple o Leaky, los resultados muy similares), este es el gráfico de error versus épocas:

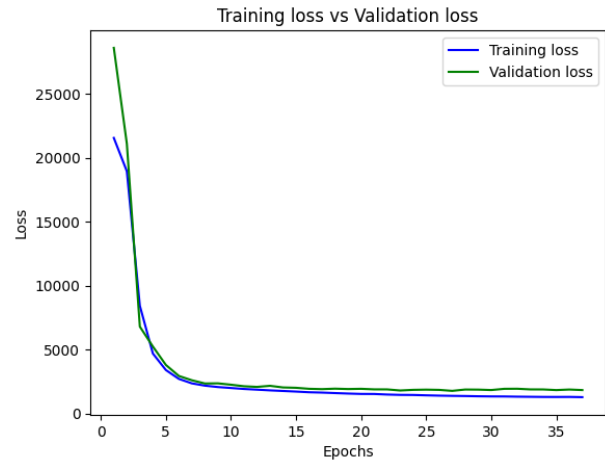


Gráfico 2:MSE Entrenamiento y Test por época.

Y este es el gráfico de predicciones versus la data real:

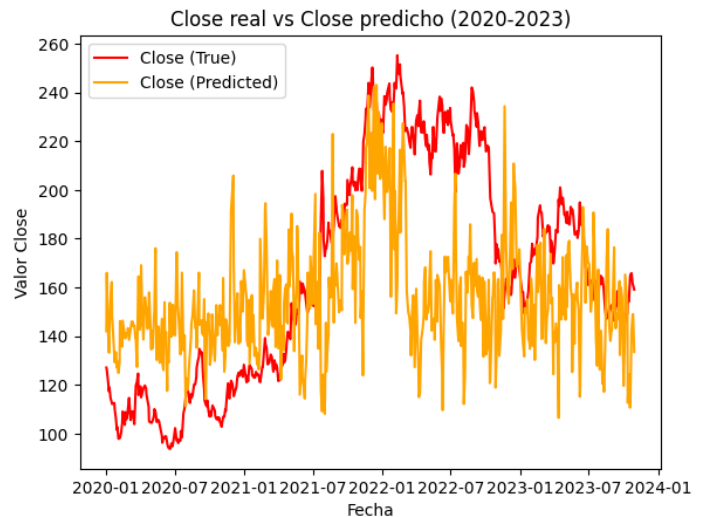


Gráfico 3: Test vs Predicciones RNN 2020-2023.

Se observa que, aunque el modelo no es capaz de explicar al 100% la variación, si hay una correlación, adicionalmente es capaz de explicar la subida que tuvo entre 2021 y 2022 debido a las heladas de 2021,

- **Árboles de Decisión:** Los resultados no fueron satisfactorios para este modelo, si bien XGBoost es capaz de predecir series de tiempo, las variables que se utilizaron (solo meteorológicas) no fueron lo suficientemente significativas en el modelo para generar un buen desempeño, tampoco fue capaz de describir el aumento de precios entre 2021 y 2022 y sus predicciones fueron casi estáticas, lo cual no genera ningún hallazgo para los caficultores y tiene un mal poder predictivo, a continuación los resultados y la gráfica de predicciones versus valor real del primer modelo.

Parámetros	MSE
1000 T, LR 0.001, SS 0.8	2539.727923
1000 T, LR 0.001, SS 0.3	2528.277565
1000 T, LR 0.001, SS 0.3, G=0.2	2529.001332

Tabla 3:MSE XGBoost Diferente Arq. Elaboración propia

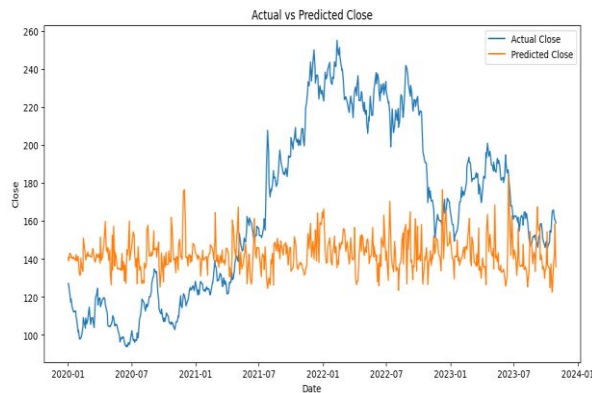


Gráfico 4: Test vs Predicciones XGBoost 2020-2023.

- **Regresión Lineal:** Los modelos de regresión lineal generaron resultados similares a los de XGBoost, solo fueron casi un 1% mejor pero nuevamente no son los resultados esperados, este tipo de modelos no es capaz de encontrar relaciones ocultas en comparación con los modelos de redes neuronales, aunque la tendencia fue menos errática que en los modelos de árboles, y aunque intenta predecir la subida de precios de 2021 y 2022, sus pronósticos siempre son demasiado optimistas, estos son los resultados y la gráfica de predicciones versus valor real del último modelo:

L	Ridge	Lasso	Elastic Net
1	2515.798	2543.114	2530.537
0.5	2518.161	2542.239	2159.393

Tabla 4:MSE Regresión Lineal, diferente regularización, elaboración propia

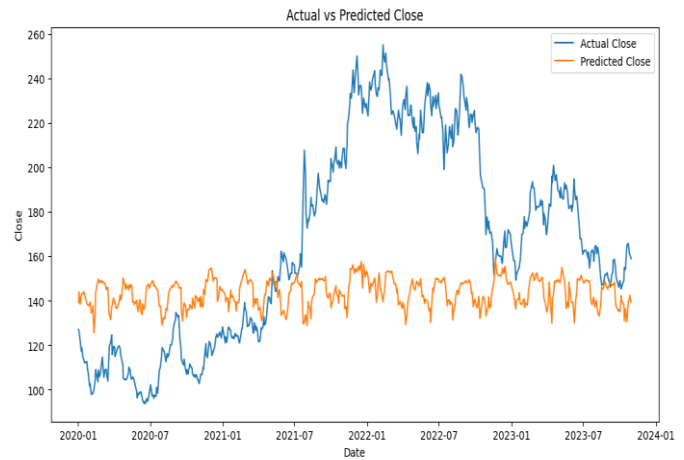


Gráfico 5: Test vs Predicciones Elastic Net, 2020-2023.

## V. DISCUSIÓN

Se observa como las redes neuronales son capaces de generar mejores pronósticos con relaciones ocultas entre los datos en comparación a otros modelos, adicionalmente su desempeño fue mucho mejor que los modelos de XGBoost o Regresión lineal (un 16% mejor en términos de MSE) para las bases de datos con retraso de 6 y 12 meses, y si bien la capacidad predictiva no es la mejor en general si es capaz de predecir el aumento de precios por factores climáticos, de manera tal que los caficultores pueden saber en los próximos 6 meses si habrá una subida de precios del café arábico si al día de hoy hay climas fríos tipo heladas en Brasil, que si bien no solo debe ser un día helado si no heladas de semanas que afecten a los cultivos de manera permanente.

Teniendo en cuenta que el retraso de 6 meses fue el mejor, se pronosticó solo para el espacio de tiempo de la subida de precios (junio 2021 a marzo 2022) obteniendo un resultado de 1824 MSE, con unas correlaciones altas para la temperatura (del -60% sobre el valor del precio) lo que indica que a medida que disminuye la temperatura el precio del café tiende a aumentar:

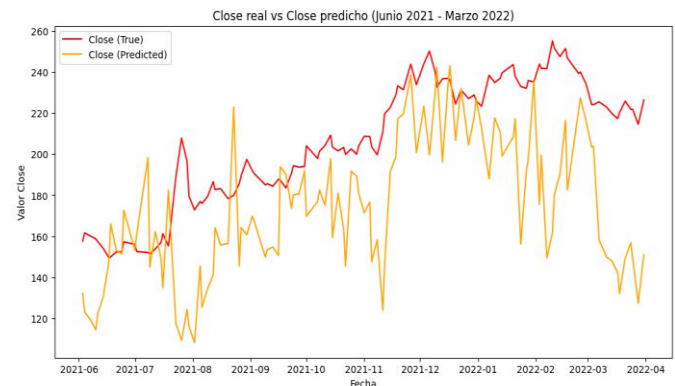


Gráfico 6: Test vs Predicciones RNN Lag 6 2021-2022.

Si bien el modelo genera unos buenos resultados, hay que mencionar que el precio del café no solo depende de su disponibilidad (lo cual cuando hay heladas tiene a reducirse)



como un efecto de la mano invisible del mercado, si no que existen otras variables las cuales no fueron tenidas en cuenta en el alcance de este estudio, como por ejemplo la situación económica global, la situación económica de los 3 países que más importan café, la apreciación o depreciación de las monedas, las políticas fiscales en estados unidos que afectan la inversión y los mercados de valores, etc.

## VI. CONCLUSIÓN

- Los caficultores colombianos pueden anticiparse a las heladas en Brasil para obtener cosechas más robustas y fructíferas. Este conocimiento les brinda una ventana de 6 meses (aunque insuficiente para sembrar y obtener nuevas cosechas) durante la cual podrían evaluar la posibilidad de almacenar lotes de café verde para su venta, siempre y cuando cuenten con la liquidez necesaria. Alternativamente, podrían invertir en el cuidado y abono de los cafetos ya sembrados, con el fin de asegurar cargas más productivas y así realizar una inversión orientada al precio futuro.
- El Retraso de 6 meses arrojó los resultados más destacados para los modelos de redes neuronales. Esto implica que, aunque la información meteorológica incide en los precios debido a la disponibilidad (ya que afecta tanto a las cosechas completas como a las de baja carga), ningún otro modelo logró prever los incrementos de precio únicamente a partir de los datos meteorológicos. De esta manera, las redes neuronales continúan cumpliendo su propósito al descubrir relaciones complejas dentro de los datos.
- Es necesario ampliar el monitoreo de los precios incorporando variables micro y macroeconómicas, tales como los límites de endeudamiento de Estados Unidos, la fortaleza de las monedas de los principales importadores de café (EE. UU., Europa/Suiza y Japón), la inflación de precios y el historial de comportamiento del mercado. Con la inclusión de estas variables, es posible que modelos como XGBoost mejoren su capacidad predictiva o alcancen niveles comparables a los de los modelos de redes neuronales.

## APLICACIÓN ESPERADA

Esperamos que la aplicación brinde a los usuarios una experiencia interactiva e informativa donde puedan ingresar datos a tiempo real de la cotización del café y clima, y, con base en los datos históricos con los que se alimentó el modelo pueda suministrar información acerca de cómo se va a comportar la cotización (Sube o baja). Esto, con el fin de que puedan tomar decisiones basadas en información relevante.

## VII. REFERENCIAS

- [1] International Coffee Organization - Historical data on the global coffee trade. (2023). [https://www.ico.org/new\\_historical.asp](https://www.ico.org/new_historical.asp)
- [2] Costa, B. (2020, January 30). Guía Concisa de las Principales Regiones Productoras de Café en Brasil. Perfect Daily Grind Español. <https://perfectdailygrind.com/es/2018/01/05/guia-concisa-de-las-principales-regiones-productoras-de-cafe-en-brasil/>
- [3] Instituto Nacional de Meteorologia - INMET. (2023.). <https://portal.inmet.gov.br/dadoshistoricos>
- [4] Coffee Statistics - Federación Nacional de Cafeteros. (2023, October 3). Federación Nacional De Cafeteros. <https://federaciondecafeteros.org/wp/coffee-statistics/?lang=en>
- [5] Ahmed, S., Brinkley, S., Smith, E., Sela, A., Theisen, M., Thibodeau, C., Warne, T., Anderson, E., Van Dusen, N., Giuliano, P., Ionescu, K. E., & Cash, S. B. (2021). Climate change and coffee quality: Systematic review on the effects of environmental and management variation on secondary metabolites and sensory attributes of *Coffea arabica* and *Coffea canephora*. *Frontiers in plant science*, 12. <https://doi.org/10.3389/fpls.2021.708013>
- [6] Gómez-Dans, J. L., et al. (2019). "Drought and Coffee in Ethiopia: Exploring the Climate Connection." *Nature Plants*, 5(5), 415-416.
- [7] INMET :: BDMEP. (2023.). <https://bdmep.inmet.gov.br/>
- [8] "Yahoo is part of the Yahoo family of brands." <https://finance.yahoo.com/quote/KC=F?p=KC=F&.tsrc=fin-srch>
- [9] International Coffee Organization (ICO). (2021). "Market Report - October 2021." Retrieved from <https://www.ico.org/documents/cy2020-21/icc-137-6e-background-marketreport.pdf>