Vigilada Mineducación

PREDICCIÓN DEL PRECIO DEL ORO EN EL MERCADO SPOT Y EL TIPO DE CAMBIO USD-COP PARA LA OPTIMIZACIÓN DEL RANGO DE COBERTURA EN DERIVADOS DE LAS COMPAÑIAS EXPORTADORAS DEL SECTOR MINERO

Prediction of the price of gold in the spot market and the USD–COP exchange rate for the optimization of the coverage range in derivatives of export companies in the mining sector

CRISTIAN ALEXANDER GALLEGO PANESSO

Tesis de grado

Asesora
Paula María Almonacid Hurtado

UNIVERSIDAD EAFIT
ESCUELA DE ECONOMÍA Y FINANZAS
MAESTRÍA EN ADMINISTRACIÓN FINANCIERA - MAF
MEDELLÍN
2024

Predicción del precio del oro en el mercado *spot* y el tipo de cambio USD–COP para la optimización del rango de cobertura en derivados de las compañías exportadoras del sector minero

Cristian Alexander Gallego Panesso
Maestría en Administración Financiera – MAF
Universidad EAFIT
cagallegop@eafit.edu.co

Resumen

En este trabajo de grado se aborda la implementación de diversos modelos de regresión de series de tiempo y *machine learning*, tales como: ARIMA, ARIMAX, SARIMA y bosques aleatorios (*Random Forest*) con el objetivo de predecir de manera precisa el precio del oro en el mercado *spot* y el tipo de cambio USD–COP. La precisión en estas predicciones es crucial para las compañías exportadoras del sector minero, ya que les permite establecer rangos de cobertura óptimos en el uso de derivados financieros. A lo largo del estudio, se evaluaron y compararon distintos algoritmos de *machine learning*, seleccionando aquellos que proporcionaron los resultados más exactos y consistentes. Los hallazgos ofrecen una herramienta valiosa para la gestión de riesgos financieros y la toma de decisiones estratégicas en el contexto de la volatilidad de los precios del oro y las fluctuaciones del tipo de cambio.

Al final del estudio se indica que el modelo ARIMAX *Rolling Forecast* aplicado en una parametrización (1,1,0) fue el modelo más acertado y consistente en el tiempo para el pronóstico del precio de ambos activos.

Palabras clave

Derivados financieros, Oro, Modelación financiera, Riesgo de Mercado, Cobertura cambiaria, bosques aleatorios

Abstract

This study addresses the implementation of various time series regression and machine learning models, such as: ARIMA, ARIMAX, SARIMA and Random Forests with the objective of accurately predicting the price of gold in the spot market and the USD–COP exchange rate. Precision in these predictions is crucial for export companies in the mining sector, as it allows them to establish optimal coverage ranges in the use of financial derivatives. Throughout the study, different machine learning algorithms were evaluated and compared, selecting those that provided the most accurate and consistent results. The findings offer a valuable tool for financial risk management and strategic decision making in the context of gold price volatility and exchange rate fluctuations.

At the end of the study, it is indicated that the ARIMAX Rolling Forecast model applied in a parameterization (1,1,0) was the most accurate and consistent model over time for the price forecasts of both assets.

Keywords: Financial derivatives, Gold, Financial modeling, Market risk, Currency hedging, Random Forest

---

Tabla de contenido

1 Introducción 4

2 Marco teórico 5
2.1 Revisión de literatura 6
2.2 Marco de referencia conceptual 6
2.2.1 Modelo Lineal Simple 7
2.2.2 Regresión Lineal Múltiple 9
2.2.3 Modelo ARIMA 10
2.2.4 Modelo SARIMAX 11
2.2.5 *Machine Learning* 11

3 Metodología **13**
3.1 Determinación de las variables de análisis 13
3.2 Transformación de las variables.....13

4 Análisis de resultados **14**
4.1 Análisis exploratorio de las variables.....19
4.2 Pruebas de estacionariedad.....19
4.3 Implementación de los modelos econométricos.....20
4.3.1 Modelo de regresión lineal múltiple 20
4.3.2 Modelo ARIMA.....22
4.3.3 Modelo ARIMAX.....24
4.3.4 Modelo SARIMA.....25
4.3.5 Modelo *Random Forest* (Bosques Aleatorios).....26

5 Resultados consolidados **29**

6 Conclusiones **30**
7 Referencias bibliográficas **32**

Índice de figuras

Figura 1. Funcionamiento de la potenciación de árboles por gradiente. 12
Figura 2. Comportamiento de las variables durante el período de tiempo estimado en el análisis.. 15
Figura 3. Comportamiento de las variables dependientes a pronosticar 16
Figura 4. Diagramas de dispersión de las variables 17
Figura 5. Mapa de correlación entre variables 18
Figura 6. Regresión lineal de los valores predichos vs valores actuales - oro 20
Figura 7. Regresión lineal de los valores predichos vs valores actuales – USD-COP 21
Figura 8. Predicción del precio del oro mediante el modelo ARIMA 22
Figura 9. Predicción del precio del USD-COP mediante el modelo ARIMA 23
Figura 10. Predicción del precio del oro mediante el modelo ARIMAX 24
Figura 11. Predicción del precio del USD-COP mediante el modelo ARIMAX 24
Figura 12. Predicción del precio del oro mediante el modelo SARIMA 25
Figura 13. Predicción del precio del USD-COP mediante el modelo SARIMA 26
Figura 14. Predicción del precio del oro mediante el modelo de *Random Forest* (XGB) 27
Figura 15. Predicción del precio del USD-COP mediante el modelo de *Random Forest* (XGB) ... 27

Índice de tablas

Tabla 1. Estadística descriptiva de las variables 14
Tabla 2. Pruebas de estacionalidad - Aumented Dickey Fuller Test 19
Tabla 3. Resultados de las métricas de rendimiento sobre las predicciones del modelo de regresión lineal múltiple para el oro 20
Tabla 4. Resultados de las métricas de rendimiento sobre las predicciones del modelo de regresión lineal múltiple para el USD-COP 21
Tabla 5. Resultados Modelo ARIMA 23
Tabla 6. Resultados modelo SARIMA 26
Tabla 7. Resultados modelo *Random Forest* oro 28
Tabla 8. Resultados modelo Random Forest USD-COP 28
Tabla 9. Resultados comparativos oro 29
Tabla 10. Resultados comparativos USD-COP 29

Índice de ecuaciones

Ecuación 1. 7
Ecuación 2 7
Ecuación 3 7
Ecuación 4 8
Ecuación 5 8
Ecuación 6 8
Ecuación 7 8
Ecuación 8 8
Ecuación 9 9
Ecuación 10 9

Ecuación 11 9
Ecuación 12 9
Ecuación 13 10
Ecuación 14 11
Ecuación 15 11

1. Introducción

El uso de los derivados financieros en las empresas colombianas de acuerdo con el Banco de la República es cada vez mayor (Corredor, 2018), el monto promedio negociado en el año 2020 se ubicó en USD 2.246 millones, alcanzando niveles récord de participación por parte de agentes locales y extranjeros debido a la pandemia, la cual desencadenó escenarios altamente volátiles para las compañías del sector minero que realizaban actividades de exportación debido a los diferentes tipos de riesgos financieros a los que se vieron expuestos a raíz de este acontecimiento macroeconómico para el que ninguna compañía estaba preparada.

Las organizaciones deben estar sujetas a efectuar sus transacciones en diferentes monedas dependiendo del mercado en el que se encuentren y de acuerdo con el país de negociación, con el fin de obtener mayores ingresos por el incremento de sus ventas; sin embargo, esta situación acarrea un riesgo cambiario, debido a que el factor de cambio no es estable y las tasas de interés de los créditos a los que acceden pueden ser altas según el momento en el que se encuentre el mercado. En el sector exportador, el problema se presenta cuando cae el precio de la divisa y la organización recibe ingresos inferiores a los esperados por la misma cantidad en ventas; esta situación pone en aprietos a las compañías cuando su planeación y su presupuesto han sido fijados con un precio de divisa puesto que proyectan sus ventas con una tasa y se encuentran con un valor inferior, lo que termina afectando sus ingresos (Fontalvo y Rodríguez, 2020).

Además de esto, las compañías comercializadoras internacionales son catalogadas como empresas altamente riesgosas para el sistema financiero, debido a la naturaleza de su negocio, y a que el sector minero es catalogado por muchos agentes gubernamentales bajo una figura de ilegalidad, debido al alto interés que tienen ciertos grupos ilegales en las diferentes comunidades en las que se extraen los metales preciosos, específicamente el oro, en el control de las diferentes minas y maquinaria que se utiliza para su extracción y que encasilla a los mineros de subsistencia y a los mineros informales dentro de estos tipos de estructuras. Esto hace que sea complejo y engorroso para estas compañías encontrar en las instituciones financieras locales tasas competitivas para realizar operaciones de cobertura cambiaria que no comprometan en gran medida sus flujos de caja.

Ante este panorama, se estudió la conveniencia de efectuar diferentes modelos de regresión de series de tiempo y *machine learning*. Se comienza evaluando los modelos de regresión lineal simple, modelo de regresión lineal múltiple, ARIMA, ARIMAX, SARIMAX y bosques

aleatorios (*Random Forest* con gradiente potenciado), con el fin de estimar un rango dinámico de cobertura y de predicción de precio tanto para el oro, como para la tasa de cambio del peso colombiano frente al dólar, para determinar a través de la metodología implementada cuál de los modelos mencionados anteriormente arroja una mayor precisión en su resultado.

Con el propósito de adentrarse en los diferentes modelos econométricos de predicción de precios de los activos en mención al inicio del documento, se da una contextualización del objetivo principal del estudio, seguido de una revisión de literatura en la que se citan algunos autores e investigaciones que abordan estudios relacionados, seguido del marco conceptual en el que se exponen las teorías de los modelos en los que se basa la investigación, el desarrollo del estudio, sus respectivos resultados, y por último, se dan las conclusiones y recomendaciones de los hallazgos.

2. Marco teórico

2.1. Revisión de literatura

En los últimos años, ha aumentado la demanda de oro de mercados emergentes como China e India, ya que el crecimiento de la clase media de estos países ha provocado un incremento de la demanda de joyas y otros productos basados en este metal. También ha aumentado la demanda de oro por parte de los bancos centrales de todo el mundo, que han incrementado sus reservas para diversificar sus activos y protegerse de las fluctuaciones monetarias (Botero, 2007).

La demanda global también se ha incrementado por la incertidumbre en los mercados financieros a raíz de los diferentes conflictos bélicos que se han venido presentando en Europa Oriental y Medio Oriente; además se ha dado un aumento de la inflación, las medidas de política monetaria contractiva en Estados Unidos y Europa, incertidumbre política a nivel local y en algunos países de la región en el que las monedas de los países emergentes han experimentado una depreciación respecto al dólar. Ante este panorama, se ha vuelto absolutamente necesaria la adopción de una política de cobertura y mitigación de riesgo por parte de las compañías que transan activos altamente volátiles como los *commodities* y en las que los resultados de sus operaciones y sus flujos de caja pueden verse altamente comprometidos por las fuertes fluctuaciones de la tasa de cambio.

Teniendo esto en cuenta, para el desarrollo del presente estudio se consultaron algunas referencias bibliográficas que abordaran la predicción de precios de diferentes activos financieros mediante la aplicación y el estudio de diferentes modelos econométricos que puedan entregar resultados altamente precisos y confiables. Dentro de estas lecturas cabe resaltar las siguientes:

Hsin, Hung – Chen, en su libro titulado "La integración de redes neuronales artificiales y la minería de datos para el pronóstico de precios futuros del oro" usa la minería de datos y las redes [1]neuronales artificiales para pronosticar los precios del oro y compararlo con el modelo autorregresivo de media móvil ARIMA. El modelo ARIMA es el modelo usado más frecuentemente para analizar datos de series de tiempo. Consiste en realizar una regresión de la variable a sus propios valores pasados y modelar el error como una combinación lineal de errores ocurriendo simultáneamente y varias veces en el pasado (Chen & Chen, 2007).

Por otra parte, Juan Manuel Candelo y Andrés Oviedo (2023) en su artículo "La volatilidad de la moneda: un análisis de la tasa de cambio colombiana y los mercados de materias primas energéticas" citan autores como (Al-Mulali y Sab, 2012; Bénassy-Quéré et al., 2007; Buetzer et al., 2016; Butt et al., 2020; Chen y Chen, 2007; Lizardo y Mollick, 2010; Narayan et al., 2008) que afirman que la causalidad de los precios de las materias primas es el factor que explica las fluctuaciones monetarias.

Con relación al uso de diversos modelos de *machine learnin*g, dentro de los cuales se encuentran los modelos de ensamble en el artículo de Zhang y Liang, (2024) "Predicción de índices de metales preciosos basada en el aprendizaje de ensamble y el método interpretable SHAP" los autores realizaron diferentes modelados y una comparación. XGB, LightGBM, CatBoost y *Random Forest*, los tres modelos con mayor desempeño predictivo, fueron seleccionados para el ensamble. Finalmente, los autores concluyen que los modelos de ensamble muestran habilidades mejoradas de predicción que los modelos individuales (Zhang & Liang, 2024).

Finalmente, Pierdzioch, C., Risse, M. (2020) en su artículo "Pronosticando retornos de metales preciosos con bosques aleatorios multivariados" utilizan el modelo de bosques aleatorios para calcular pronósticos fuera de muestra de un vector de rendimientos de cuatro metales preciosos (oro, plata, paladio y platino) y comparan los pronósticos multivariados con los univariados implícitos fuera de la muestra en bosques aleatorios ajustados de forma independiente a cada serie de retornos. Al utilizar criterios de evaluación de pronósticos univariados y multivariados llegan a la conclusión que los multivariados logran ser más precisos.

2.2. Marco de referencia conceptual

El presente estudio se basó en la elaboración de un modelo de predicción de precio del oro y del par USD–COP (peso – dólar estadounidense). El objetivo principal, estimar un rango de cobertura dinámico en la negociación de las operaciones realizadas en ambos activos financieros. Para determinar dicho rango de cobertura, se elaboraron tres modelos econométricos para cada una de las variables dependientes, con el fin de establecer cuál de estos era el modelo que mejores resultados arrojaba en el estudio, de acuerdo con los supuestos planteados inicialmente por

[1] El extracto en el que a continuación se detallan algunos conceptos de modelación y sus respectivas interpretaciones matemáticas fue extraído del trabajo de grado: Cardona y Castilla (2023) con autorización de sus autores y de la Universidad Eafit.

diferentes parámetros estadísticos.
Para el desarrollo de dichos modelos de pronóstico, se comenzó realizando una aproximación a los resultados mediante el uso de la regresión lineal simple para ambos activos con el fin de estimar si había multicolinealidad con alguna de las variables independientes analizadas y así mismo, observar el nivel de significancia de las mismas. Las regresiones usadas en el desarrollo de los primeros modelos econométricos pueden explicarse de la siguiente manera:

2.2.1 Modelo lineal simple

Ecuación 1. Modelo lineal simple

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Según James *et al.* (2021), el modelo de regresión lineal simple es una forma cuantitativa de estimar el valor de una variable dependiente representada por la variable Y, en función de una variable predictora de X, esta relación se asume que es aproximadamente lineal.
B_0 y B_1 son dos constantes desconocidas que representan el intercepto y la pendiente de la recta de regresión en el modelo lineal, y ambas constantes son conocidas dentro del modelo como los coeficientes o parámetros de la regresión.
Cuando se utilizan los datos de entrenamiento, se pueden estimar los parámetros de los coeficientes y se representan de la siguiente manera B_0 y B_1. La ecuación de regresión lineal simple se modifica generando $\hat{y}$ como el resultado en estimación de la variable independiente.

Ecuación 2

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1}\chi_i$$

La forma de estimar los parámetros de la ecuación lineal parte de minimizar la suma de errores cuadrados Q_i (*RSS*). Los errores son el reconocimiento de que la relación entre las dos variables no es lineal, y son medidos como la diferencia entre el valor real representado por Y_i y el valor estimado representado por $\widehat{y_i}$.

Ecuación 3

$$Q_i = Y_i - \widehat{y}_i$$

$$RSS = Q_1^2 + Q_2^2 + \cdots + Q_n^2$$

$$RSS = (y_1 + \beta_0 + \beta_1\chi_1)^2 + (y_2 - \hat{\beta}_0 + \hat{\beta}_1\chi_2)^2 + \cdots + (y_n - \hat{\beta}_0 + \hat{\beta}_1\chi_n)^2$$

Finalmente, mediante el uso de cálculo se estiman los valores de $\hat{B}_0$ y $\hat{B}_1$ que minimizan la suma de errores cuadrados:

Ecuación 4

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(\chi_i - \bar{\chi})(y_i - \bar{y})}{\sum_{i=1}^{n}(\chi_i - \bar{\chi})^2}$$
$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1\bar{\chi}$$

De $\hat{B}_0$ y $\hat{B}_1$ se puede estimar cual es la variabilidad de estos valores con respecto a los valores reales de β_0 y β_1 mediante la medición de los errores estándares (SE) de la estimación:

Ecuación 5

$$SE(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{\chi}^2}{\sum_{i=1}^{n}(\chi_i - \bar{\chi})^2}\right]$$
$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(\chi_i - \bar{\chi})^2}$$

Adicionalmente, la relación que existe entre la variable independiente X y la variable de- pendiente Y se debe corroborar estadísticamente. Esta relación se valida mediante el uso de pruebas de hipótesis de la siguiente manera:

Ecuación 6

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Si la prueba de hipótesis H_0 es aceptada, implica que X no se encuentra vinculada linealmente con Y, por ende, no se puede validar la relación entre las dos variables.
Validar H_0 involucra estimar qué tan alejado se encuentra el parámetro $\hat{B}_1$ de cero y dependerá de la exactitud de $\hat{B}_1$ y en la práctica se utiliza para esto un estadístico t:

Ecuación 7

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Este estadístico permite medir cuántas desviaciones estándar se encuentra alejado B_1 del cero. Con ayuda del valor P se puede aceptar o rechazar la H_0 validando o desestimando la relación lineal de las variables X y Y. Cuando el valor P es menor al nivel de significancia se rechaza H_0 , declarando de esta manera una relación entre X y Y.
Finalmente, para medir el ajuste del modelo y su capacidad de pronóstico se utilizan medidas estadísticas como el error residual estándar (RSE) y el R^2.
a raíz del error residual estándar es un estimador de la desviación estándar de los errores. En otras palabras, es el monto promedio de que la predicción se alejará del valor real y se estima de

la siguiente manera:

Ecuación 8

$$\boldsymbol{RSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n-2}}\boldsymbol{RSS}} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n-2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i}-\boldsymbol{\overline{y}_i})^{\mathbf{2}}}$$

El resultado obtenido del RSE se interpreta entonces como el número de unidades promedio que se desvía el valor estimado por el modelo frente al valor real. También se podría considerar como la falta de ajuste del modelo estimado. Su principal inconveniente radica en que, al ser una unidad de medida, no es natural saber qué tan bueno o malo puede llegar a ser este ajuste.
Por otro lado, el R^2 es otra unidad de medida de ajuste un modelo en términos de proporción, pues el resultado siempre es un valor entre 0 y 1, cuando se incluye el intercepto. Su cálculo se estima de la siguiente manera:

Ecuación 9

$$\boldsymbol{R^2} = \frac{\boldsymbol{TSS-RSS}}{\boldsymbol{TSS}} = \mathbf{1} - \frac{\boldsymbol{RSS}}{\boldsymbol{TSS}}$$

$$\boldsymbol{TSS} = \sum(\boldsymbol{y_i}-\boldsymbol{\overline{y}})^{\mathbf{2}}$$

El análisis del R^2 entonces será la porción de variación de la variable Y que puede ser explicada por la variable X.

2.2.2. Regresión lineal múltiple

Cuando una variable dependiente presenta una relación y puede ser explicada por múltiples variables independientes; es conveniente en lugar de ajustar una regresión lineal sencilla por cada variable, disponer de un modelo más robusto como es el caso de un modelo de regresión lineal múltiple, cuya representación matemática se plantea en la siguiente ecuación:

Ecuación 10

$$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1\chi_1} + \boldsymbol{\beta_2\chi_2} + \cdots\cdots + \boldsymbol{\beta_p\chi_p} + \boldsymbol{Q}$$

$$\boldsymbol{p} = \mathbf{1,2}\ \cdots\cdots\cdots\boldsymbol{n}$$

De nuevo como en el caso de la regresión lineal simple, la estimación de los coeficientes presenta un reto, y la solución para definir esta incógnita es mediante la selección de las constantes que minimizan la suma de los residuales al cuadrado RSS:

Ecuación 11

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1\chi_{i1} - \hat{\beta}_2\chi_{i2} - \cdots\cdots - \hat{\beta}_p\chi_{ip})^2$$

Con la estimación de los coeficientes que minimizan la suma de residuales al cuadrado, se evalúa igualmente la significancia de dicho coeficiente en la calidad de pronóstico general entregado por el modelo de regresión lineal múltiple, para esto nuevamente el valor P ayudará a medir la importancia relativa de la variable puntual en la estimación total.

Para medir entonces la relevancia de los coeficientes, se acude a las pruebas de hipótesis similar a como se hizo en el caso de la regresión lineal simple, en esta ocasión las hipótesis planteadas son las siguientes:

Ecuación 12

$$H_0 : \beta_1 = \beta_2 = \dots \dots = \beta_p = 0$$

$$H_1 : al\ menos\ un\ \beta_i\ es\ diferente\ de\ cero$$

Estas hipótesis son contrastadas mediante el uso de un estadístico F:

Ecuación 13

$$F = \frac{\frac{TSS - RSS}{p}}{\frac{RSS}{n - p - 1}}$$

$$TSS = \sum (y_i - \bar{y})^2$$

$$RSS = \sum (y_i - \hat{y}_i)^2$$

Si el resultado del estadístico F es superior a 1, se puede descartar la hipótesis nula y adicionalmente se puede inferir que al menos un coeficiente de todos los n asociados es diferente de cero.

Mediante el uso del valor P se puede validar la relevancia de la variable χ_i en la predicción que dicha variable puede aportar al pronóstico del modelo de regresión lineal múltiple.
Para corroborar el ajuste del modelo, de manera similar a lo definido en el modelo de Regresión Lineal Simple, se puede medir su capacidad de pronóstico mediante el uso de medidas estadísticas como el error residual estándar (RSE) y el R^2.

2.2.3. Modelo ARIMA

Como lo describe Brooks (2008), el modelo ARIMA conocido como el modelo autorregresivo integrado de promedio móvil. Es muy usado, como el modelo ARMA, en el entendimiento del comportamiento de una serie estadística y adicionalmente para la predicción de valores futuros de la serie en estudio.

Este modelo consta de tres componentes. El elemento autorregresivo (AR) relaciona el valor actual con valores pasados (retrasados). El elemento de media móvil (MA) supone que el error de regresión es una combinación lineal de errores de pronóstico pasados. Finalmente, el componente integrado (I) indica que los valores de los datos han sido reemplazados por la diferencia entre sus valores y los anteriores (y este proceso de diferenciación puede haberse realizado más de una vez).
Para la predicción, el modelo se soporta en el comportamiento de los datos en valores anteriores:

Ecuación 14

$$ARIMA(p, d, q)$$
$$p = \text{componente autorregresivo}$$
$$d = \text{componente integrada}$$
$$q = \text{componente de media móvil}$$

Ecuación 15

Matemáticamente:

$$\chi_t = -(\Delta^d \chi_t - \chi_t) + \Phi_0 + \sum_{i=1}^{p} \Phi_1 \Delta^d \chi_{t-i} - \sum_{i=1}^{p} \Theta_i Q_{t-1} + Q_t$$
$$\Phi p = \text{parámetros pertenecientes a la parte autorregresiva}$$
$$\Theta q = \text{parámetros pertenecientes a la parte de medias móviles}$$
$$Q_t = \text{término de error}$$

2.2.4. Modelo SARIMAX

SARIMAX (promedio móvil integrado autorregresivo estacional con regresores exógenos). Si bien los modelos ARIMA son bien conocidos, los modelos SARIMAX amplían el marco ARIMA al incorporar perfectamente patrones estacionales y variables exógenas.
Dentro de la biblioteca de modelos de estadísticas, la implementación de ARIMA-SARIMAX ofrece una característica valiosa: la capacidad de integrar variables exógenas como factores de pronóstico junto con la serie de tiempo principal bajo consideración. El único requisito para incluir una variable exógena es la necesidad de conocer el valor de la variable también durante el período de pronóstico. La suma de variables exógenas se realiza mediante el argumento *exog* (Amat & Escobar, 2023).

2.2.5. Machine learning

Los modelos de aumento de gradiente han ganado popularidad en la comunidad de *machine learning* debido a su capacidad para lograr excelentes resultados en una amplia gama de casos de uso, incluidas tanto la regresión como la clasificación. Aunque estos modelos han sido tradicionalmente menos comunes en la elaboración de pronósticos, pueden ser muy eficaces en este ámbito. Algunos de los beneficios clave de utilizar modelos de aumento de gradiente para realizar pronósticos son: la facilidad con la que se pueden incluir en el modelo variables exógenas, además de variables autorregresivas, la capacidad de capturar relaciones no lineales entre variables, la alta escalabilidad, que permite que los modelos manejen grandes volúmenes de datos. Algunas implementaciones permiten la inclusión de variables categóricas sin necesidad de codificación adicional (Amat & Escobar, 2023).

El presente estudio concluye abarcando la aplicación de un modelo de bosques aleatorios con gradiente XGB (Extreme Gradient Boosting) para series de tiempo, con el fin de

determinar la precisión de pronóstico del mismo, comparado con los modelos de regresión planteados anteriormente.

Cuando se utiliza el aumento de gradiente para la regresión, los alumnos más débiles son los árboles de regresión, y cada árbol de regresión asigna un punto de datos de entrada a una de sus hojas que contiene una puntuación continua. XGB minimiza una función de objetivo regularizada (L1 y L2) que combina una función de pérdida convexa (según la diferencia entre las salidas de destino y previstas) y un plazo de penalización para la complejidad de modelos (es decir, las funciones de árboles de regresión). La capacitación avanza de forma iterativa, agregando nuevos árboles que predicen los residuos de errores de los árboles anteriores que se combinan después con los árboles anteriores para realizar la predicción final. Se denomina potenciación del gradiente porque utiliza un algoritmo de gradiente descendente para minimizar la pérdida cuando se agregan nuevos modelos.
Esta es una descripción gráfica sobre el funcionamiento de la potenciación de árboles por gradiente.

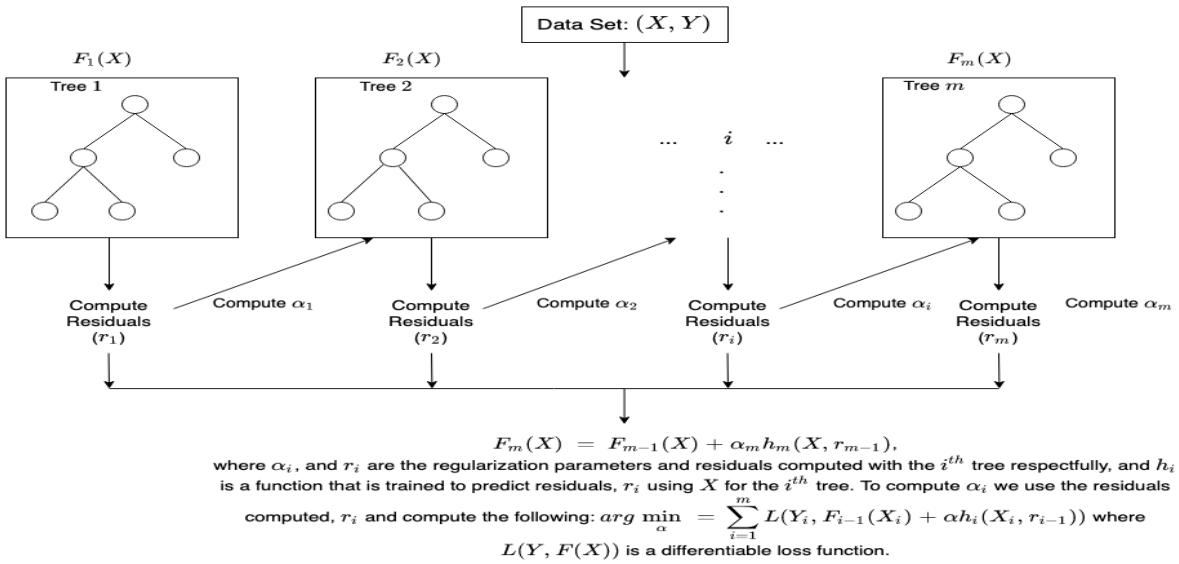

Figura 1. Funcionamiento de la potenciación de árboles por gradiente.

Fuente: Amazon SageMaker Documentation. (2024). Potenciación de árboles por gradiente [Figura]. Funcionamiento de XGBoost - Amazon SageMaker

3. Metodología

3.1. Determinación de las variables de análisis

Teniendo como propósito de este estudio la predicción de un rango dinámico de cobertura basado en los precios de cotización de la onza de oro y la tasa de cambio del par USD-COP, como apoyo también de la revisión bibliográfica citada en el presente estudio y de acuerdo al criterio del autor del mismo de acuerdo por su experiencia en el trading en los mercados financieros, se evaluaron y eligieron las siguientes variables, las cuales tienen una relación directa en el comportamiento de las dos variables dependientes citadas anteriormente:

- Precio diario cierre de la onza de oro.
- Precio diario de cierre del par USD-COP.
- Precio diario de cierre del índice DXY (dólar).
- Precio diario de cierre del S&P500.
- Precio diario de cierre del petróleo referencia WTI.
- Precio diario de cierre de los Tesoros de Estados Unidos a 10 años.
- Precio diario de cierre del índice VIX.

Para el desarrollo de los modelos econométricos se tomó una base de datos con un total de 17.795 datos, los cuales abarcan los precios diarios de los activos mencionados desde el 1° de enero de 2013, hasta el 26 de abril del año 2024. Una vez analizados los datos, se procede a normalizar las series con el fin de transformar la escala de distribución de los mismos y así poder realizar una comparativa de las distintas variables.

3.2.Transformación de las variables

Para la realización del presente estudio se debió realizar una estandarización de las fechas de los diferentes índices de estructuración para su revisión, lo que sirvió como una llave para combinar todos los conjuntos de datos de todas las variables y conservar únicamente el precio de cierre de las mismas en días iguales.

Para la imputación de los valores faltantes (empty values), se realizó una interpolación lineal en dirección hacia adelante, en el caso del petróleo de referencia Brent que se descartó al inicio del estudio por problemas de multicolinealidad se hallaron 79 datos faltantes por lo que también se descartó la implementación de la variable en el modelo de datos. Finalmente, para resolver los problemas de escala se procedió a realizar una normalización de datos que resultara en que todos los valores terminaran en la misma escala. Para esto, se tomaron los mínimos y máximos de los conjuntos de datos a través de la transformación MinMaxScaler, el cual escala y traduce cada característica individualmente y le otorga valores entre 0 y 1 a un conjunto de entrenamiento.

Matemáticamente, la transformación está dada por la expresión:

$$X\_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))$$
$$X\_scaled = X\_std * (max - min) + min$$

Como base, se toma el modelo de cinco años, comprendido entre las fechas desde el 01 de enero de 2019 y el 26 de abril de 2024, ya que fue el período en el que se encontró mayor cantidad de datos disponibles sin necesidad de imputar mayores series de tiempo. De igual manera, se estimó que las series de corto plazo tenían mayores problemas de tendencia, por lo que se desestimó usar fechas posteriores a un año.

4. Análisis de resultados

4.1. Análisis de las variables

Con el fin de identificar diferentes patrones, tendencias o valores significativos se realiza un análisis descriptivo de las variables para determinar si hay diferencias significativas en los datos que puedan afectar la proyección, y si es del caso, aplicar las transformaciones necesarias.

Tabla 1. Estadística descriptiva de las variables

Estadístico	wti	usd_cop	10y_yield	sp500	vix	usd_index	gold_y
count	3.324	3.325	3.311	33.240	33.240	33.250	33.250
mean	66,6239	3.232,3616	2,3666	3.000,0325	17,8394	95,0931	1.516,4190
std	22,7915	787,4670	0,9107	995,9039	7,0876	7,3119	302,1600
min	7,7900	1.759,5000	0,5120	1.457,2	9,1400	79,0900	1.051,7400
25%	50,3075	2.868,9100	1,7555	2.093,9000	13,1900	92,1400	1.259,5466
50%	64,2400	3.186,5000	2,3260	2.793,8500	15,7750	96,0500	1.366,0500
75%	82,8600	3.833,2800	2,8570	3.949	20,7125	99,3800	1.805,2000
max	128,2600	5.104	4,9900	5.254,400	82,6900	114,1100	2.390,4500

Para complementar el análisis de las variables seleccionadas, se graficó el comportamiento de cada una de ellas en el horizonte de tiempo estimado inicialmente tal y como se muestra en la figura 2.

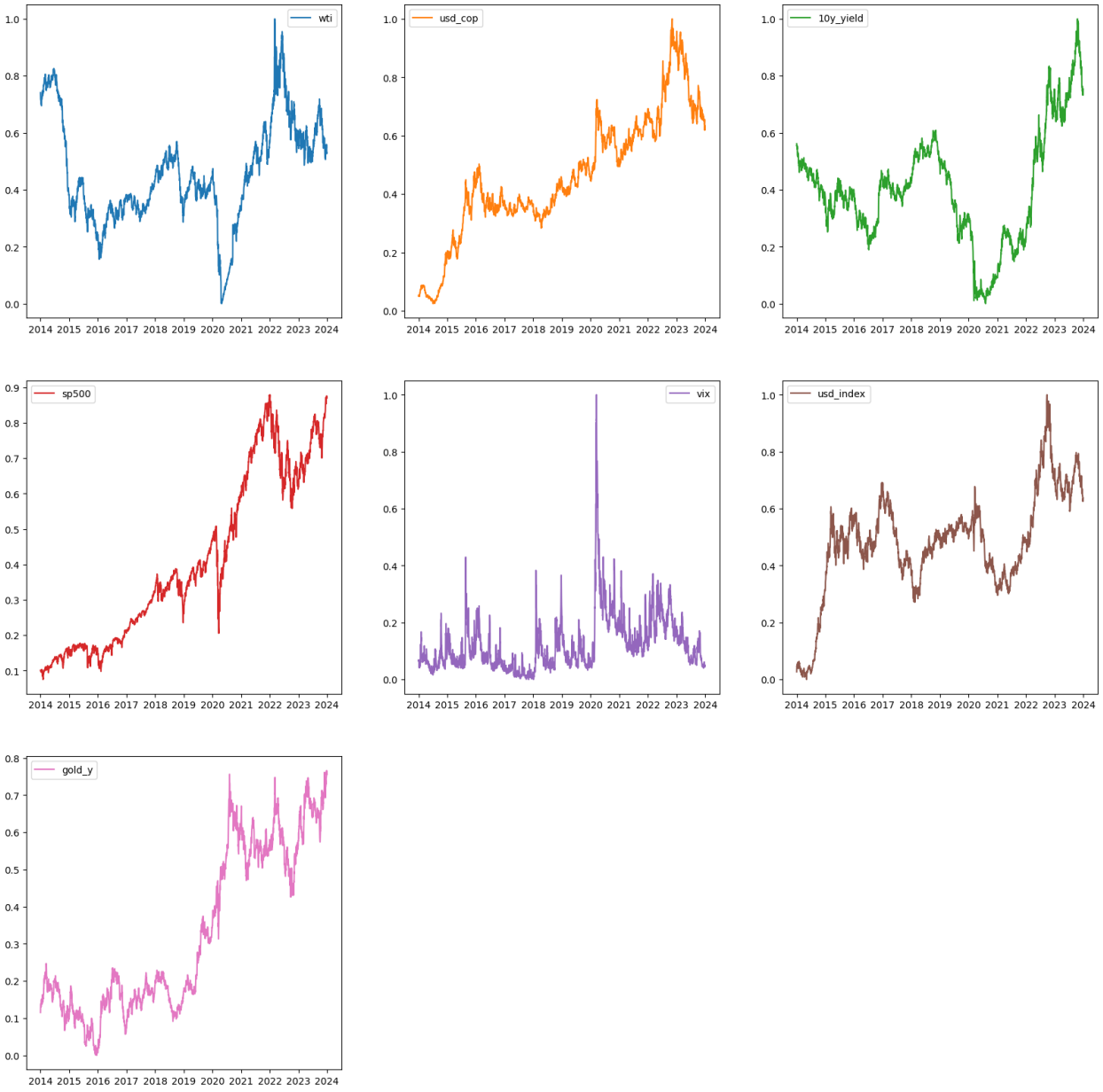

Figura 2. Comportamiento de las variables durante el período de tiempo estimado en el análisis

Figura 3. Comportamiento de las variables dependientes a pronosticar

Como puede observarse en la figura 3, el oro y el dólar históricamente han tenido una correlación negativa, debido a que el oro se considera como un activo refugio cuando hay incertidumbre en los mercados financieros.

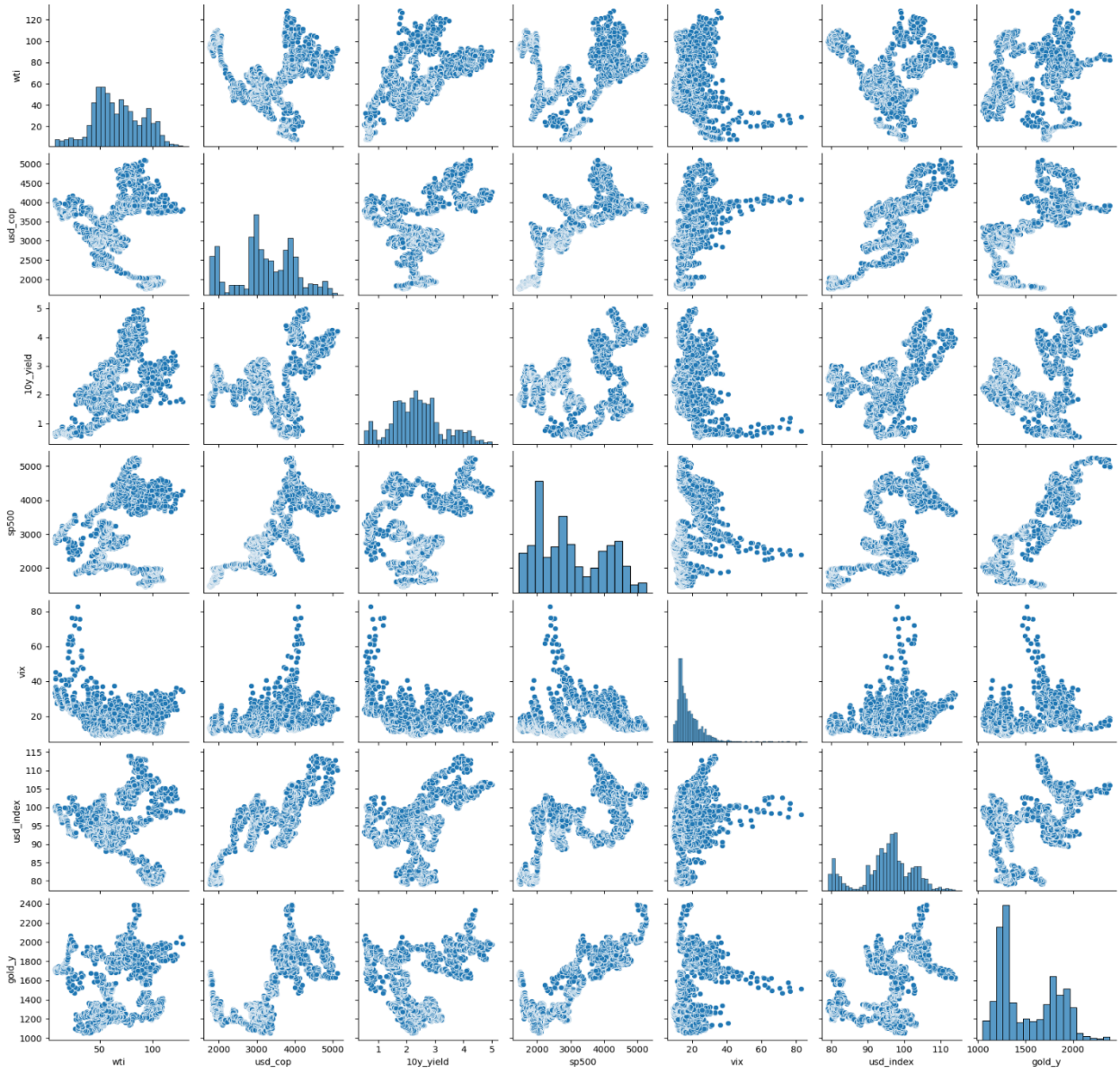

Figura 4. Diagramas de dispersión de las variables

En el diagrama de dispersión de cada una de las variables podemos observar la correlación lineal de cada una de las variables, donde se resalta la alta correlación entre el USD-COP y el USD index y la correlación inversa entre el oro y los bonos del Tesoro estadounidense a 10 años, de lo cual podemos concluir en qué medida pueden interferir la cotización de cada uno de estos activos para el pronóstico de las dos variables dependientes a predecir, lo que podremos ver expresado numéricamente en la figura 5 del mapa de correlación expuesto a continuación:

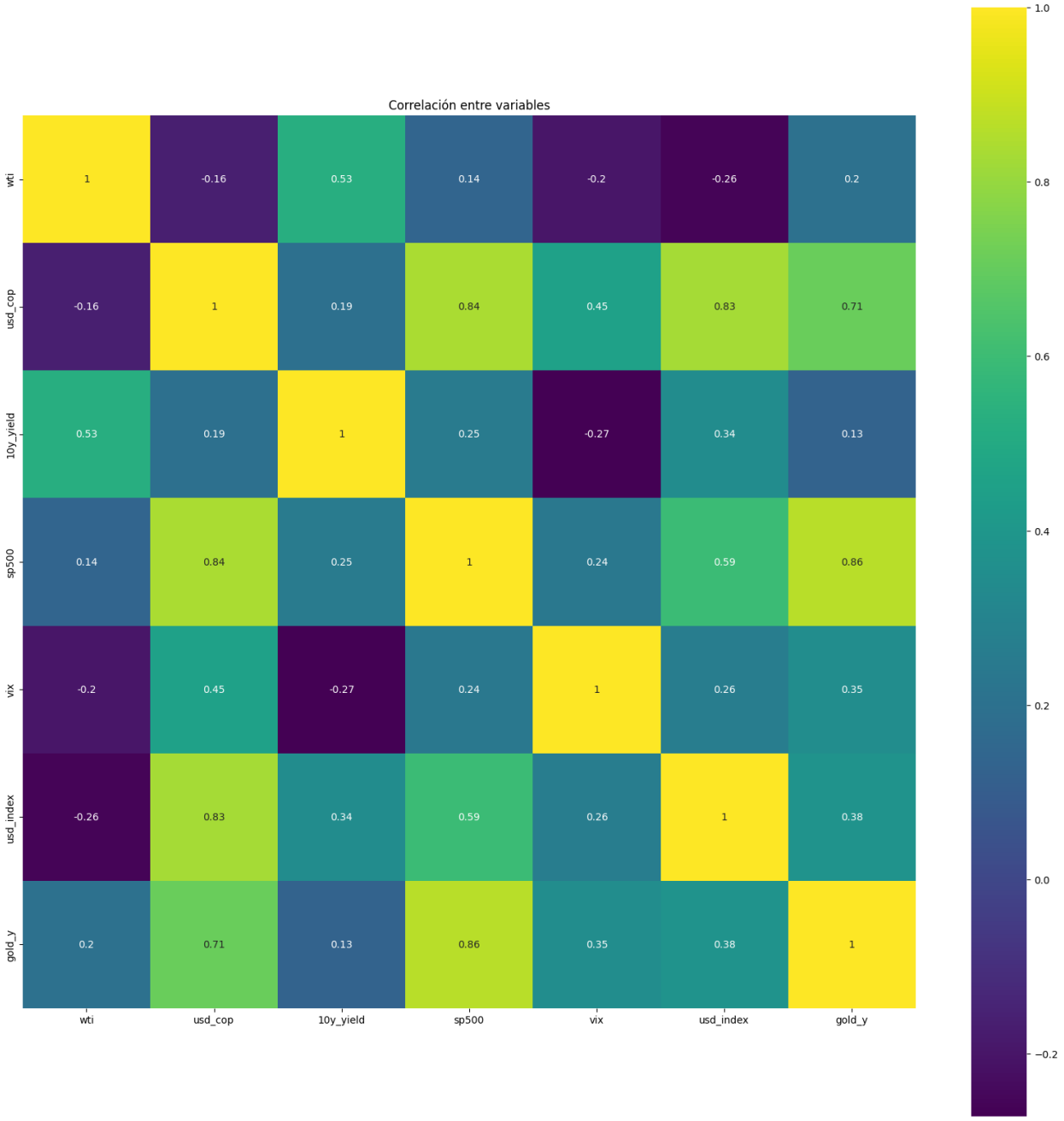

Figura 5. Mapa de correlación entre variables

Con el fin de evitar presentar problemas de multicolinealidad entre las variables, se tomaron en cuenta para el análisis dos variables adicionales, las cuales fueron el petróleo de referencia Brent y el par de divisas EUR/USD. Pero debido a que estas tenían una alta correlación con algunas de las variables independientes del estudio, se optó por seleccionar las variables con la mayor significancia entre sí, las cuales fueron el petróleo WTI y el USD Index con el propósito de obtener mejores resultados en las predicciones de los modelos posteriores al evitar un aumento del R cuadrado.

4.2. Pruebas de estacionalidad

Tabla 2. Pruebas de estacionalidad - Aumented Dickey Fuller Test

Serie histórica	Dickey Fuller P-Value
WTI	0.5703
USDCOP	0.3711
10Y Yield	0.8790
S&P 500	0.5921
VIX	0.0008
USD Index	0.5349
Gold	0.3444

De acuerdo con los resultados de la tabla anterior, los p values resultantes nos permiten descartar la hipótesis nula del test al indicarnos que las variables independientes no son estacionarias, con excepción de la variable VIX, pero su estacionariedad no afecta los resultados del pronóstico.

4.3. Desarrollo de los modelos econométricos de series de tiempo

4.3.1. Modelo de regresión lineal múltiple

Para el ejercicio se comenzó realizando un modelo de regresión lineal múltiple para ambas variables, En el caso del oro esta regresión se realizó a un año con el fin de estimar cuál era el valor R cuadrado y la significancia del error cuadrático medio RMSE para este período de tiempo, obteniendo el resultado que se muestra en la tabla 3.

Tabla 3. Resultados de las métricas de rendimiento sobre las predicciones del modelo de regresión lineal múltiple para el oro

Métrica de rendimiento	Resultado
Coeficiente de determinación "R^2"	0.761
RMSE	0.0054

De acuerdo con el resultado obtenido, la métrica R cuadrado nos indica que el 76,1% de la variabilidad en el precio del oro es explicada por las variables incluidas en el modelo.

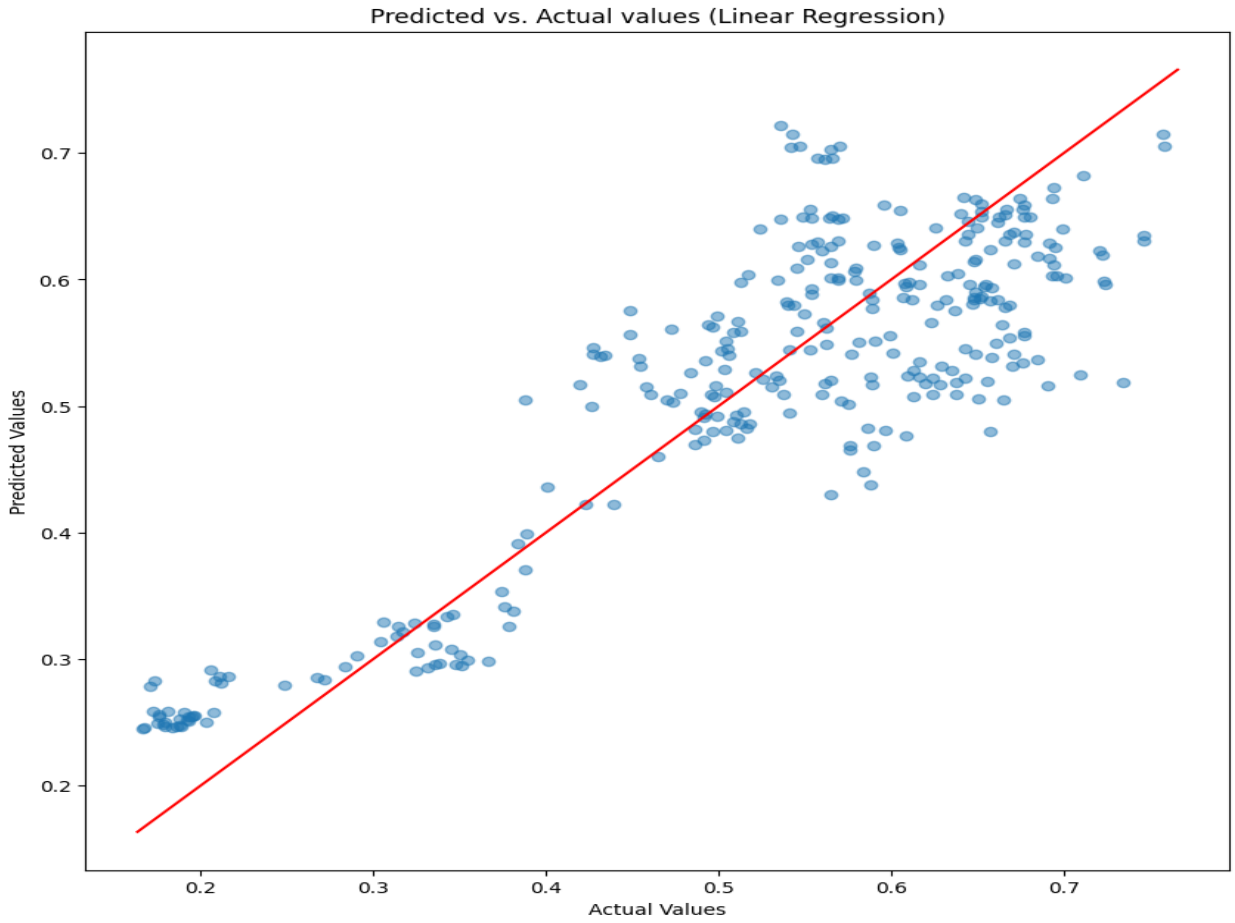

Figura 6. Regresión lineal de los valores predichos vs. valores actuales - oro

Tal como evidencia la figura anterior, los valores se van alejando de la media a medida que va aumentando la periodicidad a estimar, por esto podemos inferir que la regresión lineal múltiple para el oro es mucho más fiable si se realiza en intervalos de tiempo de corto plazo.

Tabla 4. Resultados de las métricas de rendimiento sobre las predicciones del modelo de regresión lineal múltiple para el USD-COP

Métrica de rendimiento	Resultado
Coeficiente de determinación "R^2"	0.7283
RMSE	0.0051

Como podemos observar en la tabla anterior, al igual que en la regresión lineal con el oro, para el par USD-COP también tenemos un coeficiente de correlación bastante alto y un RMSE muy bajo, por lo que podemos asumir que la regresión lineal también muestra una aproximación bastante cercana del pronóstico a los datos reales evaluados, como lo podremos observar en la gráfica 7:

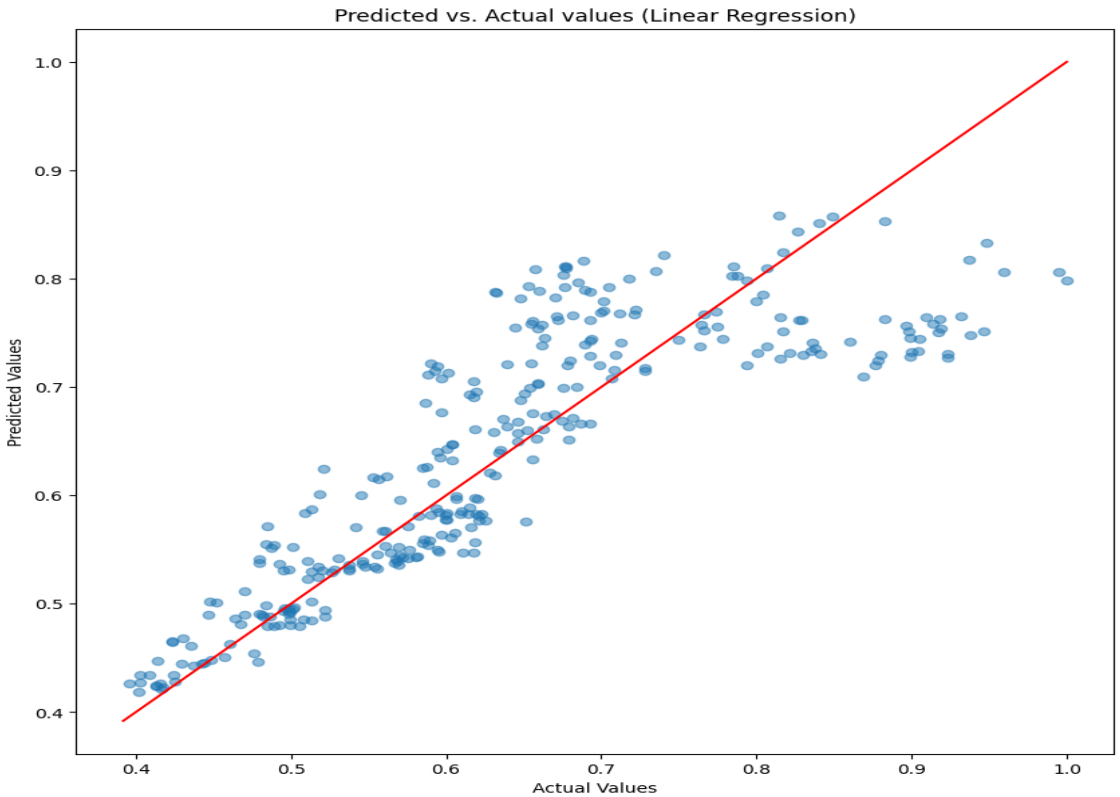

Figura 7. Regresión lineal de los valores predichos vs. valores actuales – USD-COP

Del gráfico anterior podemos concluir, que al contrario de lo que pasa en el gráfico de regresión lineal del oro en donde los valores se concentran mucho más al final de la serie de tiempo, para el USD/COP es todo lo contrario, ya que, si bien también se muestra una dispersión a la media a medida que se va ampliando la periodicidad del estudio, los valores se concentran mucho más y con mayor precisión en rangos de tiempo de muy corto plazo.

4.3.2. Modelo ARIMA

Para el presente estudio del pronóstico se utilizó una serie ARIMA [1,1,0] ya que es el modelo que mejor se adapta al pronóstico y arroja mejores métricas como se puede observar a continuación:

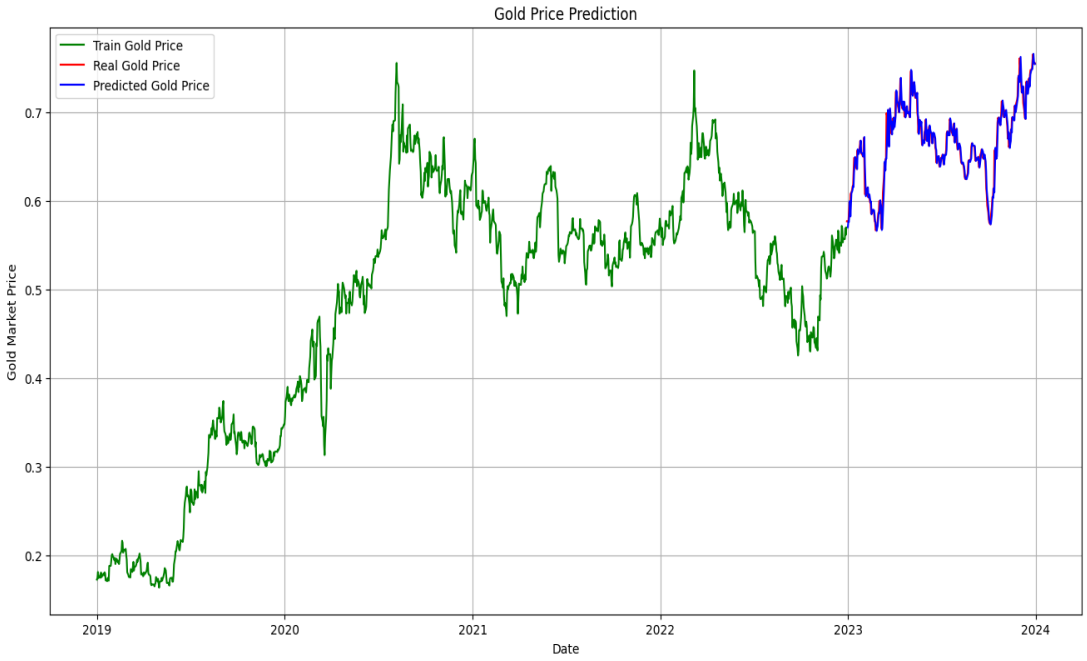

Figura 8. Predicción del precio del oro mediante el modelo ARIMA

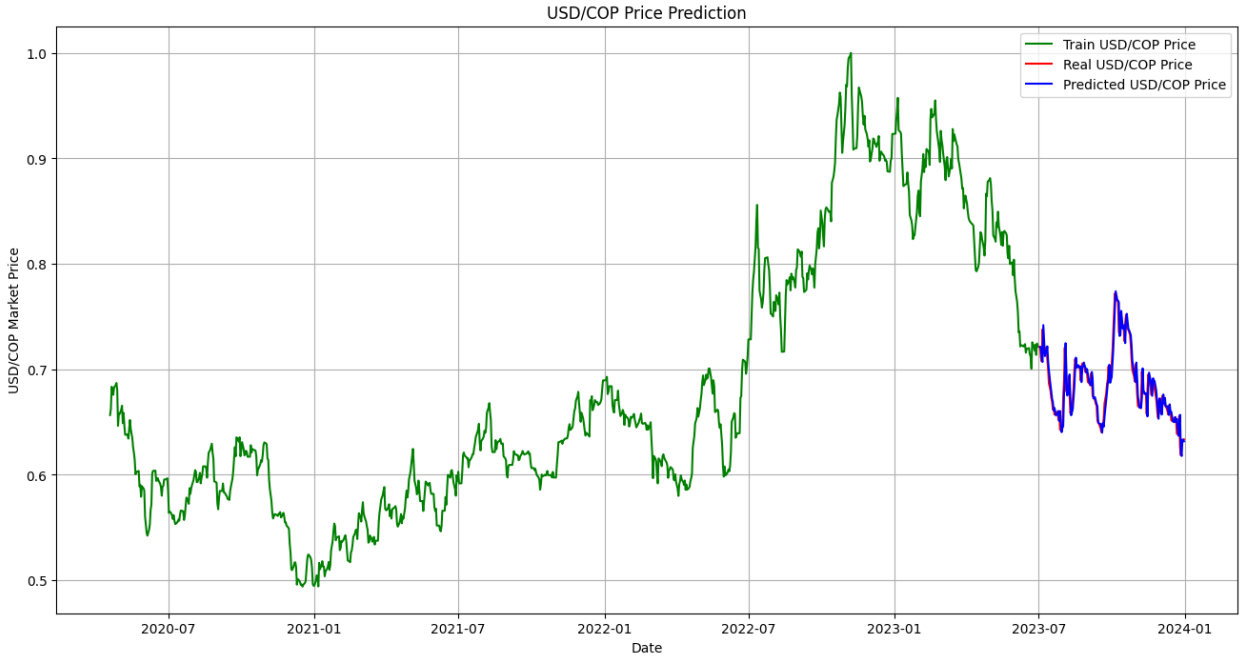

Figura 9. Predicción del precio del USD-COP mediante el modelo ARIMA

Tabla 5. Resultados modelo ARIMA

Activo	Modelo	AIC	BIC
GOLD	ARIMA (1,1,0)	-9766.102	-9755.397
USD/COP	ARIMA (1,1,0)	-10130.810	-10120.105

4.3.3. Modelo ARIMAX

Debido a que la mayoría de los datos utilizados en el modelo ARIMA son datos no estacionarios, se consideró pertinente realizar una estacionalización de los mismos, e incluir las variables exógenas, con el fin de determinar el comportamiento real del pronóstico con la inclusión de dichas variables y su precisión.

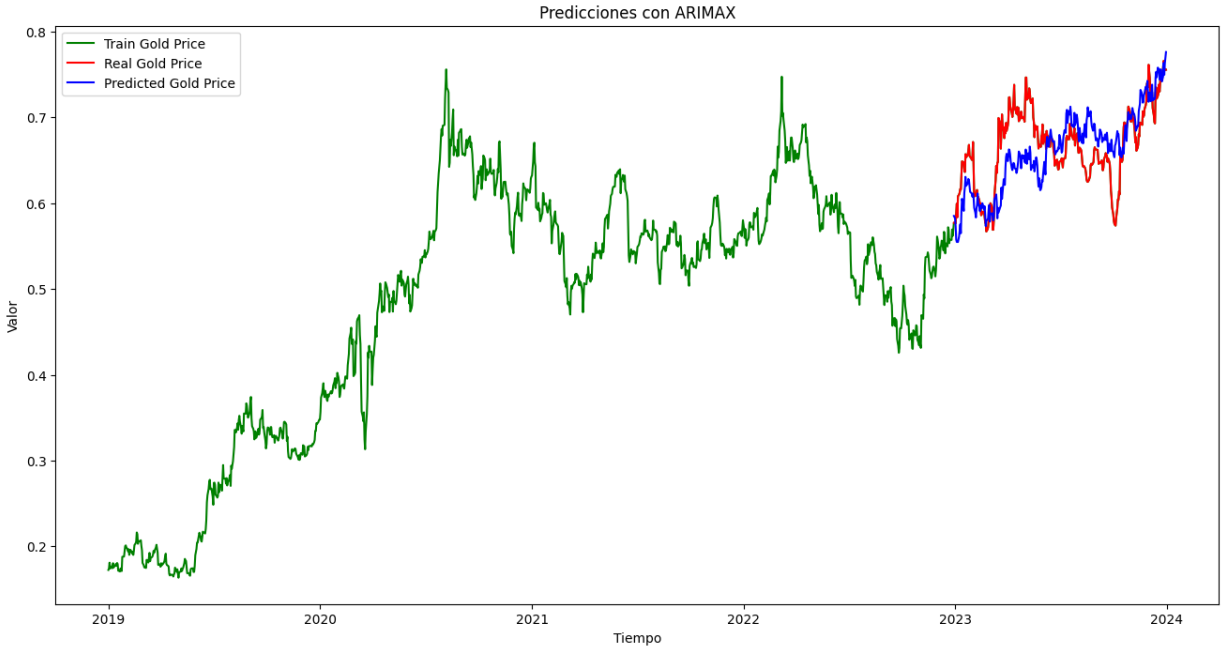

Figura 10. Predicción del precio del oro mediante el modelo ARIMAX

Figura 11. Predicción del precio del USD-COP mediante el modelo ARIMAX

4.3.4. Modelo SARIMA

Con el fin de complementar los modelos de series de tiempo analizados, se decidió usar el modelo SARIMA, ya que refleja la característica de variación estacional en series de tiempo.

El modelo SARIMA desarrollado, si bien es un modelo adecuado para el análisis de series de tiempo, no es tan preciso para el modelado de plazos más amplios, debido a que las series de datos se vuelven muy aleatorias entre sí, especialmente con activos tan variables como por ejemplo la tasa de cambio.

Para realizar un *fit* (ajuste) adecuando del modelo, se debió reducir el porcentaje de datos de prueba de 20% al 10% y se debió aumentar el porcentaje de datos de entrenamiento de 80% a 90%.

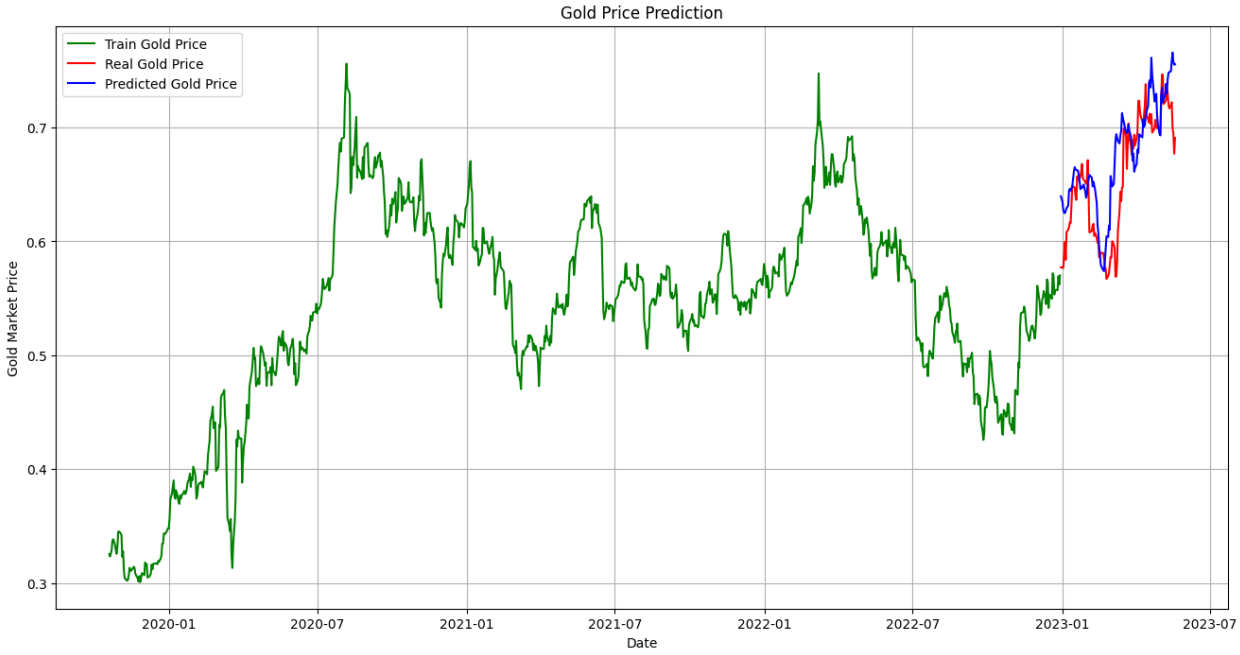

Figura 12. Predicción del precio del oro mediante el modelo SARIMA

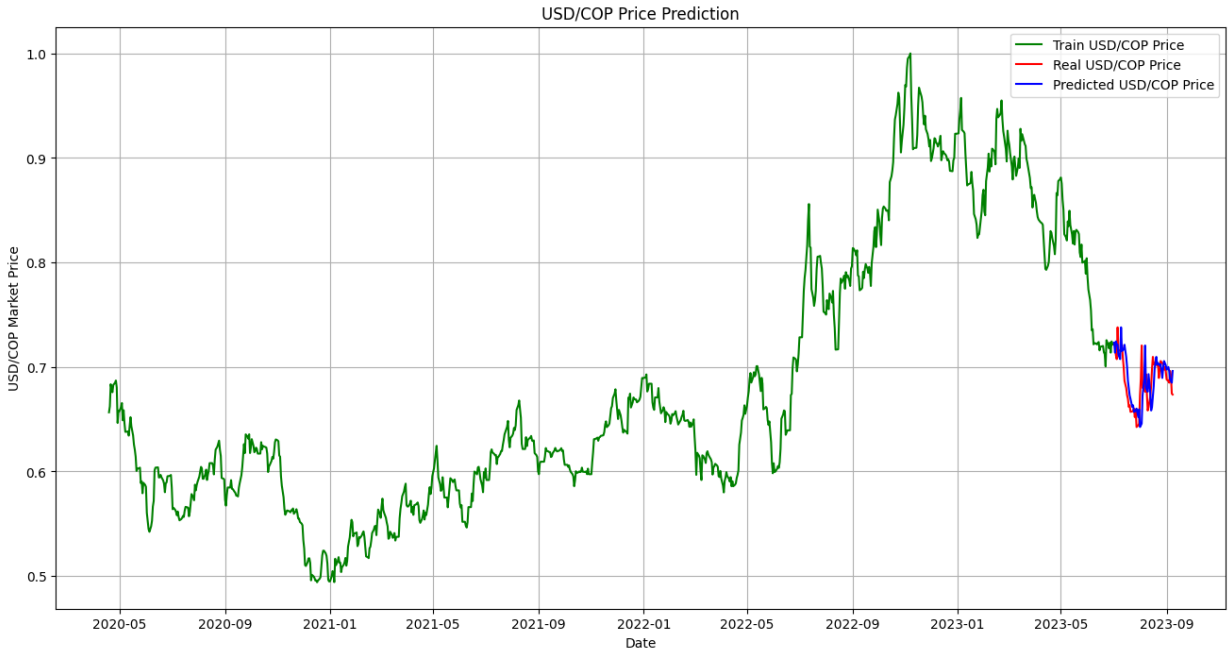

Figura 13. Predicción del precio del USD-COP mediante el modelo SARIMA

Tabla 6. Resultados modelo SARIMA

Activo	Modelo	AIC	BIC
GOLD	SARIMA (0,1,0,120)	-2257.872	-2252.598
USD/COP	SARIMA (1,1,0)	-1124.020	-1918.774

4.3.5. Machine learning

Finalmente, para la implementación del modelo de bosques aleatorios o *Random Forest* con gradiente potencializado (XGB) se toman todas las variables disponibles para un período de tiempo de cinco años y se identifica que, si bien la serie de tiempo captura con buena precisión el nivel de pronóstico, no lo hace de la misma manera con las fluctuaciones tal y como se puede ver en las siguientes figuras:

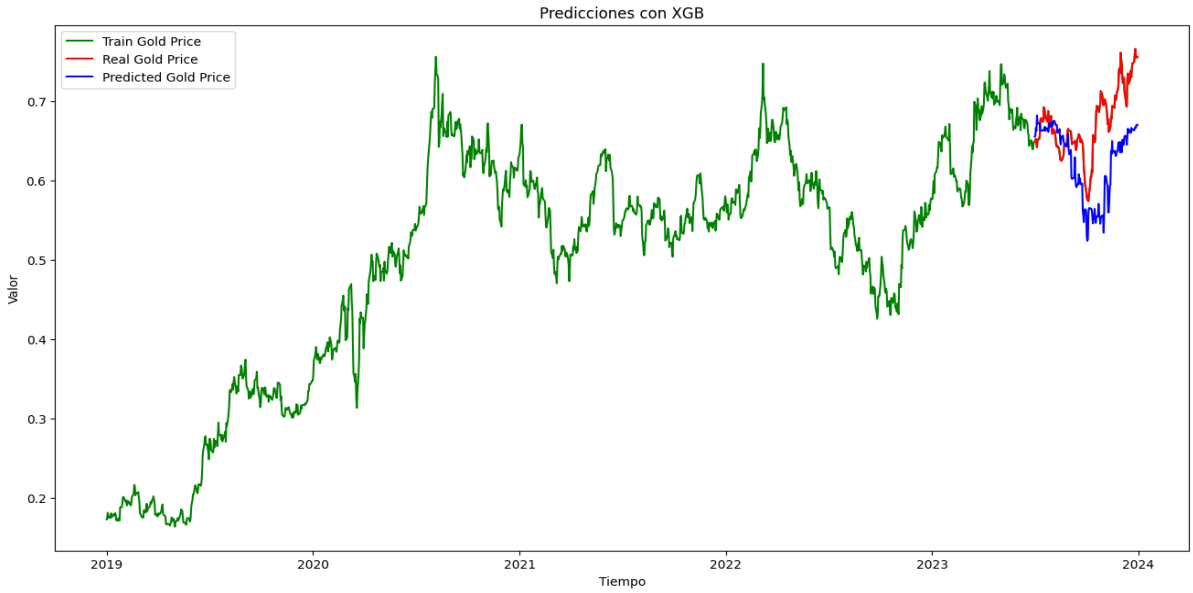

Figura 14. Predicción del precio del oro mediante el modelo de *Random Forest* (XGB)

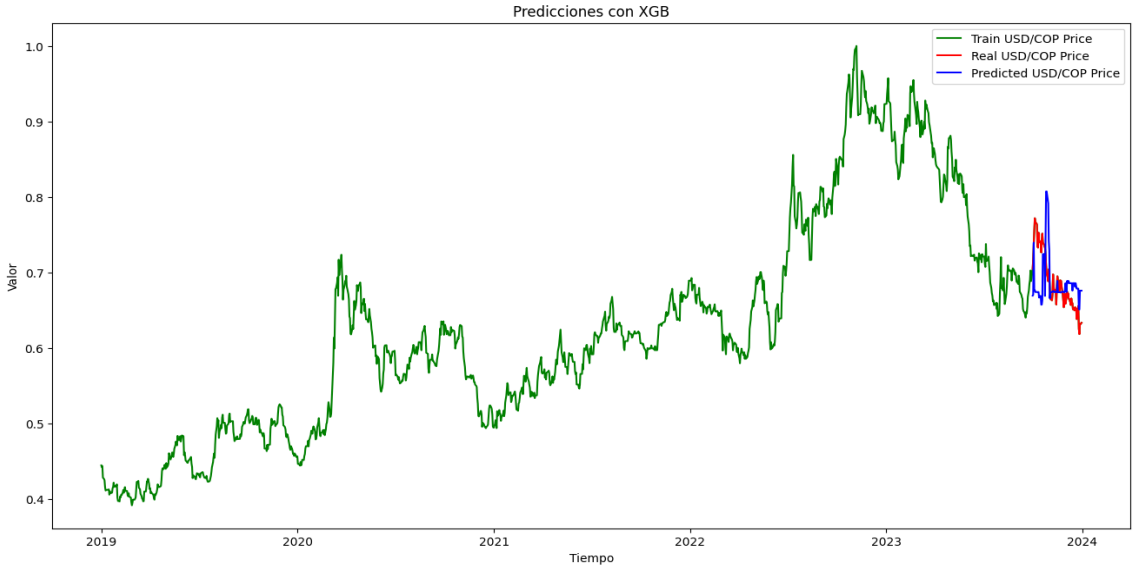

Figura 15. Predicción del precio del USD-COP mediante el modelo de *Random Forest* (XGB)

Tabla 7. Resultados modelo *Random Forest* oro

Métrica de rendimiento	Resultado
RSME (0) [0]	0.15053
RMSE (0) [1204]	0.00549
RSME (1) [0]	0.17872
RSME (1) [1204]	0.06834

Tabla 8: Resultados modelo *Random Forest* USD-COP

Métrica de rendimiento	Resultado
RSME (0) [0]	0.14004
RMSE (0) [627]	0.00686
RSME (1) [0]	0.07392
RSME (1) [627]	0.05258

En las tablas anteriores podemos observar el entrenamiento del modelo en el cual se realizaron dos configuraciones (0,1), también podemos ver los resultados de cada una de las iteraciones y destacar que en cada uno de los modelos el error medio cuadrático (RSME) va cambiando, haciéndose cada vez más bajo, es decir más preciso.

5. Resultados consolidados

En el presente apartado se muestra una tabla de resumen comparativa que se elaboró a partir de los resultados obtenidos en las parametrizaciones de cada uno de los modelos desarrollados para cada una de las variables dependientes, con sus respectivas métricas y resultados.

Tabla 8. Resultados comparativos oro

Y: Precio Oro
X: USD-COP, USD Index, Vix, Bond Yield (10 years), WTI

	Modelo	Regresión lineal múltiple (OLS)	ARIMAX (Rolling Forecast)	SARIMA	SARIMAX	*Random Forest* (XGBoost)
Métrica	RSME	0,0054	0,0001	0,0014	0,0016	0,0046
	Log Likelihood	1824	4885	1130	3388	N/A
	AIC	-3633	-9766	-2258	-6772	N/A
	BIC	-3595	-9755	-2253	-6762	N/A
	Jarque-Bera (JB)	6,60300	2145,5900	1,0800	737,6200	N/A
	Prob (JB)	0,04	0,00	0,58	0,00	N/A

Tabla 9. Resultados comparativos USD-COP

Y: USD-COP
X: Precio Oro, USD Index, Vix, Bond Yield (10 years), WTI

Regresión lineal múltiple (OLS)	ARIMAX (Rolling Forecast)	SARIMA	SARIMAX	*Random Forest (XGBoost)*
0,0039	0,0001	0,0003	0,0022	0,0021
1985	5067	963	3608	N/A
-3956	-10131	-1924	-7212	N/A
-3919	-10120	-1919	-7201	N/A
133,0060	742,8500	2,5700	273,3500	N/A
0,00	0,00	0,28	0,00	N/A

6. Conclusiones

El presente estudio tuvo el objetivo de analizar los modelos más implementados para el pronóstico de series de tiempo y así determinar cuál era el modelo más preciso para el pronóstico del oro y de la tasa de cambio colombiana, por lo que a través del análisis se obtuvieron las siguientes conclusiones:

La regresión lineal múltiple no es el modelo más adecuado para el pronóstico de oro y dólar-peso, debido a que sus resultados mostraron que a pesar de tener un R cuadrado alto para ambos modelos (más del 70%), no tiene la misma precisión en el modelado para horizontes de tiempo con un plazo mayor a un año, ya que los pronósticos se comienzan a alejar de la media al no capturar las variaciones de precio de corto plazo en finanzas.

El modelo ARIMA (1,1,0) mostró una gran precisión en términos de pronóstico para ambos activos, pero sin tener en cuenta las variables independientes, que en gran medida explican el comportamiento tanto del oro como del dólar debido a su correlación con cada una de estas variables, por lo que a pesar de tener resultados muy exactos, no son del todo fiables para determinar el propósito del estudio.

El modelo ARIMAX, no capturó el nivel de los movimientos de ambos activos de una manera tan precisa como el ARIMA, pero sí captura muy bien la tendencia, incorporando las variables exógenas y además de esto, fue el modelo que mostró mejores resultados en términos de rendimiento y en todas sus métricas.

Los modelos SARIMAX capturaron en mayor medida la realidad de las series de tiempo a predecir (valores más altos para Log-Likelihood, y más bajos para AIC, BIC), por lo que también se puede determinar que el modelo es una buena alternativa para la predicción del precio de los activos, ya que muestra una consistencia en términos de desempeño.

Para el modelado con Random Forest, se determinaron los niveles de significancia de cada una de las variables para el pronóstico y se realizó una predicción univariable en la que se determinó cada una de las variables a predecir como una variable dependiente e independiente al mismo tiempo. Si bien el modelo mostró un resultado aceptable en términos de precisión, el error cuadrático obtenido fue de los más altos en comparación con los demás modelos.

Finalmente, como conclusión acerca de la implementación de modelos de Machine Learning para la predicción del oro y el dólar, se recomienda explorar nuevas variables que no tengan una correlación tan alta o baja con las variables dependientes como en el caso del oro, Dollar Index y VIX respectivamente, con el fin de reemplazarlas y observar cómo se comporta el modelo bajo la introducción de un nuevo set de variables e hiperparámetros.

En conclusión, los modelos de predicción de precio contribuyen en gran medida a la realización de las coberturas cambiarias en las compañías mineras, debido a que estas se exponen a grandes riesgos asociados a la tasa de cambio y de mercado debido a la alta volatilidad que tienen los precios del metal precioso por su alta sensibilidad a los eventos geopolíticos y económicos, sumado a la variación constante del peso colombiano frente al dólar, que actualmente se encuentra entre uno de los pares de monedas con mayor volatilidad en el mundo, lo que en términos de riesgos se convierte en una doble exposición. Por ende, al tener las compañías importadoras y exportadoras del sector minero un pronóstico, lo más cercano posible a la realidad sobre los precios de ambos activos, podrían evitar pérdidas innecesarias en sus flujos de caja a raíz de la toma de decisiones imprecisas en la monetización de los flujos en dólares recibidos o por pagar, al igual que en la cobertura de sus posiciones en oro a través del mercado de futuros.

De igual manera, teniendo en cuenta que la cobertura frente la tasa de cambio se realiza mediante derivados financieros como futuros, forwards y swaps, esto beneficiaría la eficiencia en su negociación al poderse anticipar el precio en el mercado spot como la tasa futura, lo que se traduciría en una eficiencia operativa y de costos por devaluación.

7. Referencias

1. Amat R, J. & Escobar Ortiz, J. (2023). *Skforecast* (Version 0.11.0) [Computer software]. https://doi.org/10.5281/zenodo.8382787
2. Botero, M. M. (2007). *La ruta del oro*. Fondo Editorial Universidad EAFIT.
3. Brooks, C. (2008) Introducción a la econometría para las finanzas. Cambridge University Press.
4. Candelo-Viáfara, J. M. & Oviedo-Gómez, A. (2023). La volatilidad de la moneda: un análisis de la tasa de cambio colombiana y los mercados de materias primas energéticas. *Cuadernos de Economía*, *42*(89), 177-201.
5. Cardona Restrepo J. & Castilla Rueda. R. A. (2023). Predicción del precio de transacción sobre el tipo de cambio XAU-USD (Oro) para el mercado de contado del commoditie a corto plazo, Medellín, Colombia. *Universidad Eafit.*
6. Chen, S. & Chen, H. (2007). Oil prices and real exchange rates. *Energy Economics 29*(3),390-404. https://doi.org/10.1016/j.eneco.2006.08.003

7. Corredor, A. (2018). El uso de forwards peso dólar en las empresas colombianas del sector real. *Borradores de Economía,* 1058. http://repositorio.banrep.gov.co/bitstream/handle/20.500.12134/9524/be_1058.pdf?sequence=8&isAllowed=y
8. Fontalvo Jaramillo, K. &, Rodríguez Velásquez, P. G. (2020). Uso y aplicación de derivados financieros en empresas colombianas: Forwards y swaps, Bogotá, Colombia. Universidad Eafit. https://repository.eafit.edu.co/bitstream/handle/10784/26500/KatherinePaola_FontalvoJaramillo_PedroGiovanny_RodriguezVelasquez_2020.pdf?sequence=2&isAllowed=y

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021) An Introduction to Statistical Learning: With Applications in R. 2nd Edition. *Springer*. https://doi.org/10.1007/978-1-0716-1418-1.

9. Pierdzioch, C. & Risse, M. (2020). Forecasting Precious Metal Returns with Multivariate Random Forests. *Empirical Economics.* Volume 58, Issue 3, 1167-1184.
10. Zhang, Y & Liang, M. O, H. (2024). Predicition of Precious Metal Index Based on Ensemble Learning and SHAP Interpretable Method, Computational Economics. *BanRep Cultural.* https://www.banrep.gov.co/sites/default/files/publicaciones/archivos/be_860.pdf