



**Comparación de modelos de series temporales
ARIMA, SARIMAX y LSTM para la predicción del índice COLCAP**

DAVID SANTIAGO OSORIO ARISTIZÁBAL

Proyecto de Grado

Asesor

Édison Valencia Díaz

UNIVERSIDAD EAFIT
ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA
MAESTRÍA CIENCIAS DE LOS DATOS Y ANALÍTICA

MEDELLÍN

2024

CONTENIDO

0. INTRODUCCIÓN	4
1. PLANTEAMIENTO DEL PROBLEMA	5
3. OBJETIVOS	7
3.1. GENERAL	7
3.2. ESPECÍFICOS	7
4. MARCO TEÓRICO	8
5. DISEÑO METODOLÓGICO	10
6. RESULTADOS	12
7. CONCLUSIONES	29
8. ANEXO NOTEBOOK	30
9. REFERENCIAS	30

Resumen

Este trabajo de grado realiza un análisis comparativo de tres enfoques de modelado predictivo para el índice COLCAP: ARIMA, SARIMAX y LSTM. El objetivo es evaluar la precisión y eficacia de cada modelo para proporcionar predicciones precisas del índice bursátil utilizando variables exógenas.

Se inicia implementando el modelo ARIMA debido a su simplicidad y, posteriormente, se utiliza el modelo SARIMAX, que extiende el ARIMA incorporando componentes estacionales y variables exógenas, el cual permitió identificar las variables Financial Times Stock Exchange de Colombia (FTSE COLOMBIA), índice de Mercados Emergentes (EMM), precio del petróleo crudo en su especie brent (BRENT) y el Global X MSCI Colombia ETF (GXG) como las de mejor relación con el COLCAP. Finalmente se implementó un modelo de red neuronal recurrente LSTM (Long Short-Term Memory) con las variables exógenas obtenidas en el modelo SARIMAX. Con los resultados del análisis comparativo, se concluyó que el modelo SARIMAX tiene buena capacidad para ajustarse a la serie COLCAP pero su capacidad de predicción se ve limitada al conocimiento futuro de las variables exógenas. Además, el modelo LSTM permite superar esta limitación al aprender las relaciones entre las variables exógenas y la serie temporal objetivo.

Palabras clave: Series Temporales, COLCAP, ARIMA, SARIMAX, LSTM, Variables Exógenas, Redes Neuronales Recurrentes.

0. INTRODUCCIÓN

El índice COLCAP es un referente crucial para el mercado bursátil colombiano, ya que refleja el comportamiento de las acciones más líquidas de la Bolsa de Valores de Colombia. Entender el comportamiento de este índice es de suma importancia para inversionistas y analistas financieros pues proporciona información valiosa para la toma de decisiones económicas y financieras. Teniendo esto en cuenta, la capacidad de anticipar las fluctuaciones del COLCAP puede contribuir significativamente al desarrollo de estrategias de inversión.

El origen de este estudio radica en la necesidad de explorar y comparar diferentes enfoques de modelado predictivo que puedan registrar de manera efectiva las dinámicas del índice COLCAP. Para este propósito, se han seleccionado tres modelos representativos en el ámbito de las series temporales y el aprendizaje automático: ARIMA, SARIMAX y LSTM. Cada uno de estos modelos ofrece distintas ventajas y desafíos, por lo que una comparación de estos modelos resulta esencial para identificar el enfoque más adecuado.

El alcance de este trabajo abarca desde la implementación de modelos tradicionales de series temporales, como el ARIMA, hasta un modelo de redes neuronales recurrentes, específicamente el LSTM. Se implementa también el modelo SARIMAX, el cual permite agregar variables exógenas y componentes estacionales, ampliando así la capacidad predictiva del modelo. Las variables exógenas consideradas en este estudio incluyen EMBI, DOLAR (USDCOP), CIB, FTSE COLOMBIA, EMM, BRENT, GXG, ILF y TASA DE INTERÉS, seleccionadas por su relevancia y correlación con el COLCAP, las cuáles serán detalladas en el desarrollo del trabajo.

La metodología empleada comprende las siguientes etapas: comprensión del problema y de los datos, preparación de los datos, análisis descriptivo, análisis de estacionariedad, implementación de los modelos predictivos, evaluación de su desempeño por medio de métricas como el Error Cuadrático Medio (MSE), el Error Absoluto Medio (MAE) y el Error Porcentual Absoluto Medio (MAPE), comparación de los resultados obtenidos, y finalmente las conclusiones.

El impacto de este estudio para el área de interés radica en la provisión de una guía detallada y comparativa sobre la eficacia de distintos modelos predictivos aplicados al índice COLCAP.

1. PLANTEAMIENTO DEL PROBLEMA

El índice financiero COLCAP representa el mercado accionario colombiano, por lo que entender su comportamiento es un desafío importante para diferentes grupos de interés, como fondos de inversión, analistas financieros y otros tipos de inversores. Este entendimiento es crucial, ya que permite a estos grupos tomar decisiones informadas sobre inversiones, gestión de riesgos y estrategias financieras. Para comprender este comportamiento, es común que se utilicen modelos tradicionales de series de tiempo, como ARIMA y SARIMAX, que han demostrado ser efectivos para registrar patrones y tendencias en datos históricos. Sin embargo, con el avance de la inteligencia artificial y el aprendizaje automático, han surgido modelos de redes neuronales como las LSTM, que ofrecen una alternativa para una predicción más precisa de las series temporales. Estos modelos son capaces de aprender dependencias a largo plazo y manejar relaciones no lineales y complejas en los datos, lo que los hace especialmente adecuados para los mercados financieros dinámicos y volátiles.

Al comparar estos enfoques, se podrá identificar el más adecuado para diferentes contextos y necesidades de predicción. Además, el análisis resultante será más profundo, ya que no es necesario limitarse a un solo modelo. La combinación de los resultados de todos puede ofrecer una perspectiva más completa y robusta, mejorando la precisión y confiabilidad de las predicciones. Este enfoque multimodelo puede ayudar a registrar diferentes aspectos de las dinámicas del mercado que un solo modelo podría pasar por alto, proporcionando así una herramienta más poderosa para la toma de decisiones en el mercado bursátil colombiano.

2. JUSTIFICACIÓN

Potenciar la capacidad para predecir activos financieros como el índice COLCAP, considerando un mayor número de variables relacionadas y diferentes modelos, es relevante para la diversidad de agentes interesados en el desempeño del índice, incluyendo inversionistas, gestores de portafolios, empresas multinacionales, instituciones reguladoras y gubernamentales. La implementación de un enfoque basado en técnicas analíticas, puede mejorar la precisión de la predicción en contraste con enfoques tradicionales, y de esta forma conducir a una mejor toma de decisiones en diferentes ámbitos de interés de los agentes. Al incrementar la exactitud y confiabilidad de las predicciones, se facilita la identificación de oportunidades de inversión, la gestión de riesgos y el diseño de políticas económicas más efectivas, beneficiando así a todos los sectores involucrados.

3. OBJETIVOS

3.1. GENERAL

Evaluar y comparar la precisión predictiva de los modelos ARIMA, SARIMAX y LSTM para la predicción del índice COLCAP, incorporando variables exógenas relevantes, con el fin de identificar el modelo más adecuado para la optimización de estrategias de inversión y la toma de decisiones informadas en el mercado bursátil colombiano.

3.2. ESPECÍFICOS

- Desarrollar modelos ARIMA, SARIMAX y LSTM utilizando datos históricos del índice COLCAP, empleando variables exógenas en los modelos SARIMAX y LSTM, que permitan mejor desempeño.
- Evaluar el desempeño predictivo de los modelos estudiados utilizando métricas acordes a cada uno, principalmente el error cuadrático medio (MSE).
- Comparar los resultados obtenidos en cada uno de los modelos e identificar las mejores variables exógenas que se adaptan al modelo.

4. MARCO TEÓRICO

Modelos de Predicción en el Mercado Bursátil

La predicción del precio de los índices bursátiles ha sido una tarea desafiante para las personas que trabajan en el campo financiero y otras escalas relacionadas debido a las características de volatilidad y ruido (Gao et al., 2020). En este contexto, es crucial comparar diferentes enfoques de modelado predictivo para identificar el más adecuado.

Para este propósito, los modelos ARIMA (Autoregressive Integrated Moving Average) son ampliamente utilizados para el análisis y la predicción de series temporales debido a su capacidad para capturar patrones de autocorrelación en datos univariados. Sin embargo, ARIMA es un modelo lineal y, por lo tanto, puede tener limitaciones al manejar relaciones no lineales y complejas en los datos financieros. La comparación entre modelos ARIMA y el modelo LSTM (Long Short-Term Memory), es importante para evaluar el desempeño predictivo entre un modelo lineal y uno no lineal porque ayuda a contrastar el rendimiento predictivo de los modelos tradicionales de series de tiempo frente a modelos de aprendizaje profundo más avanzados (Bukhari et al., 2020). Esta comparación permite a los investigadores evaluar la eficacia de las redes LSTM para capturar dependencias a largo plazo y patrones no lineales en datos financieros, los cuales ARIMA puede no modelar adecuadamente (Bhandari et al., 2022).

LSTM puede adaptarse mejor a cambios en los datos y patrones emergentes, lo que lo hace más versátil para ajustarse a la naturaleza dinámica y no estacionaria de los mercados financieros en comparación con ARIMA, que es más estático en su enfoque de modelado (Bukhari et al., 2020).

Redes Neuronales Recurrentes

Las redes neuronales recurrentes, especialmente las LSTM han surgido como alternativas poderosas para la predicción de series temporales en los mercados financieros. Las redes LSTM se destacan en la predicción de índices financieros, como el S&P 500, gracias a su efectividad para aprender dependencias temporales de largo plazo, utilizando celdas de memoria para superar el problema de los gradientes que desaparecen en otros tipos de redes neuronales. Estas características hacen que LSTM sea más adecuado para modelar la volatilidad y el comportamiento no lineal de los precios de las acciones en comparación con modelos de tipo ARIMA (Bhandari et al., 2022).

LSTM se usa comúnmente para predecir los precios de las acciones basándose en datos históricos, incorporando varios indicadores y factores del mercado para pronosticar tendencias futuras con precisión. (Hochreiter et al., 1997). Además, los modelos LSTM son eficaces para predecir métricas financieras como precios de cierre, volúmenes de negociación y movimientos de índices, lo que ayuda en la toma de decisiones y la gestión de riesgos en el sector financiero (Bhandari et al., 2022).

Variables Exógenas en la Predicción de Series Temporales

El mercado bursátil colombiano es un componente esencial para la economía del país y el índice COLCAP juega un papel crucial al reflejar el comportamiento de las acciones más líquidas de la Bolsa de Valores de Colombia. Este índice es calculado por la firma proveedora de índices MSCI y se compone de los 20 emisores y las 25 acciones más líquidas del mercado (*Bolsa de Valores de Colombia / BVC*, 2024). Comprender y predecir el comportamiento del COLCAP es fundamental para diferentes grupos de interés, incluyendo fondos de inversión, analistas financieros y otros inversores.

Las variables exógenas pueden influir en los movimientos ascendentes y descendentes del índice, lo que puede afectar la toma de decisiones de los inversores. Al analizar y clarificar el impacto de estas variables, se puede obtener una mejor comprensión de la dinámica del mercado y tomar decisiones más informadas en relación con el riesgo y la volatilidad (Mascio et al., 2023).

Algunas de las variables exógenas que pueden afectar las predicciones de la serie del índice COLCAP son:

- **El Emerging Markets Bond Index (EMBI)** es un índice calculado por JP Morgan Chase, que se calcula como la diferencia entre la tasa de interés de los bonos soberanos de los mercados emergentes y los bonos del Tesoro de Estados Unidos. Es una medida para evaluar el riesgo de que un país no pague las obligaciones a los acreedores (Diaz Tagle et al., 2008).
- **El DÓLAR** es la tasa de cambio representativa del mercado (TRM) en Colombia, que se calcula con base en las operaciones de compra y venta de divisas entre intermediarios financieros en el mercado cambiario colombiano (Banco De La República, 2024).
- **CIB** es el nemotécnico de la acción de Bancolombia tipo ADR en Estados Unidos, que representa 4 acciones preferenciales de Bancolombia en el mercado colombiano.

GRUPO BANCOLOMBIA en el mercado colombiano es la acción con más peso en el índice COLCAP (MSCI COLCAP Index Methodology, 2021).

- **El FTSE Colombia Index** está compuesto por grandes y medianas empresas colombianas que forman parte del FTSE All-World Index, incluyendo también empresas de pequeña capitalización (FTSE Colombia Indices, 2024).
- **El ETF iShares MSCI Emerging Markets (EMM)** busca replicar los resultados de inversión de un índice compuesto por valores de renta variable de alta y mediana capitalización de mercados emergentes, como el colombiano (BLACKROCK, 2024).
- La variable **BRENT** corresponde al precio del petróleo que marca referencia en mercados europeos.
- **El Global X MSCI Colombia ETF (GXG)** invierte en las mayores y más líquidas acciones colombianas (GLOBALXETFS, 2024).

Métricas de Evaluación de Modelos Predictivos

Las métricas de evaluación comúnmente utilizadas en los trabajos de analítica incluyen MAPE (Mean Absolute Percentage Error), RMSE (Root Mean Squared Error) y MSE (Mean Squared Error) como indicadores del rendimiento, teniendo en cuenta que se busca que ambas medidas funcionen mejor cuanto más bajos sean los valores obtenidos (Vijh et al., 2020).

5. DISEÑO METODOLÓGICO

El desarrollo del proyecto sobre la modelación y predicción del índice COLCAP sigue un enfoque basado en técnicas de análisis de las series temporales, de esta forma:

- **Comprensión del problema y de los datos:** se describe el contexto de la necesidad de analizar la serie de tiempo correspondiente al índice COLCAP, la cual se aborda en las secciones de problema y justificación. Se identifican cuáles son los datos con los que se cuenta y se busca identificar cuáles de estas son las más relevantes para el análisis propuesto.
- **Preparación de los datos:** la base de datos se construye a partir de fuentes públicas como el portal del Banco de la República, el portal web del DANE e Investing. Las series exógenas se alinean temporalmente con la serie diaria del COLCAP y se ajustan a las mismas fechas. Las observaciones se registran de lunes a viernes, incluyendo los días festivos en Colombia. En caso de faltar datos en una fecha debido a un festivo, se imputa el dato correspondiente al día hábil anterior.
- **Análisis descriptivo:** se realiza un análisis descriptivo de las series temporales, calculando estadísticas básicas como media, mediana y desviación estándar. Se incluye también un análisis de correlación entre las variables. Además, se visualizan las series temporales para entender patrones básicos y tendencias a lo largo del tiempo.
- **Análisis de estacionariedad:** se llevan a cabo pruebas de estacionariedad como la prueba de Dickey-Fuller aumentada (ADF) para determinar la necesidad de transformaciones adicionales en las series temporales, como la diferenciación.
- **Modelado:** inicialmente, se ajusta un modelo ARIMA a la serie temporal del COLCAP desde un enfoque univariado. Posteriormente, a partir de los resultados obtenidos, se ajusta un modelo SARIMAX que incluye componentes de estacionalidad y variables exógenas. Finalmente, se implementa un modelo de red neuronal recurrente, específicamente una LSTM (Long Short-Term Memory), utilizando las variables exógenas. Los hiperparámetros de la LSTM se ajustan para optimizar el rendimiento del modelo.
- **Evaluación de Modelos y Comparación:** Los resultados se evalúan utilizando la métrica del Error Cuadrático Medio (MSE) para medir la calidad del ajuste de cada modelo. Además, se analiza la capacidad predictiva de los modelos a través de gráficos que comparan las predicciones de cada modelo en contraste con los valores reales del índice COLCAP.
- **Conclusiones:** finalmente, se presentan las conclusiones del estudio, destacando los hallazgos principales y las implicaciones de los resultados obtenidos.

6. RESULTADOS

6.1 Análisis descriptivo

Los datos utilizados en este estudio corresponden a las siguientes series temporales: EMBI, DOLAR (USDCOP), CIB, FTSE COLOMBIA, EMM, BRENT, GXG, ILF y TASA DE INTERÉS, cuyos datos fueron recolectados desde el 10 de febrero del 2009 hasta el 30 de abril del 2024 en temporalidad diaria para cada una de las series y para los días laborales. En caso de que la fecha corresponda a un día festivo en Colombia, se completa el dato con el valor del día hábil anterior.

Estadísticos básicos

Para el análisis descriptivo se calcularon estadísticas básicas para cada serie temporal, incluyendo la media, mediana, desviación estándar, valores mínimo y máximo. A continuación, se presenta un resumen de dichas estadísticas:

	COLCAP	EMBI	DOLAR	CIB	FTSE COLOMBIA	EMM	BRENT	GXG	ILF	TASA_INTERES_PM
count	3971.000000	3971.000000	3971.000000	3971.000000	3971.000000	3971.000000	3971.000000	3971.000000	3971.000000	3971.000000
mean	1441.325467	2.344530	2866.270672	43.717033	4354.844984	41.283289	77.558514	48.190375	33.772073	5.243264
std	216.872950	0.846866	887.572650	13.036328	672.569856	5.423707	24.308511	22.619155	8.613220	2.931062
min	825.170000	0.960000	1747.650000	16.230000	2289.550000	19.820000	19.330000	9.640000	15.640000	1.750000
25%	1291.675000	1.740000	1928.625000	32.620000	3881.800000	38.505000	57.905000	30.130000	27.070000	3.500000
50%	1440.580000	2.060000	2922.880000	42.950000	4351.910000	40.960000	75.510000	39.080000	32.060000	4.500000
75%	1614.050000	2.860000	3616.000000	55.370000	4904.915000	43.980000	101.115000	72.440000	40.090000	5.750000
max	1942.370000	5.220000	5117.000000	70.490000	5807.490000	57.960000	127.980000	95.380000	54.940000	13.250000

Figura 1. Estadística descriptiva. Fuente: Elaboración propia

Visualización de las Series Temporales

Se realizaron los gráficos de línea para cada serie temporal, lo que permitió observar la evolución de los valores de cada variable a lo largo del tiempo.

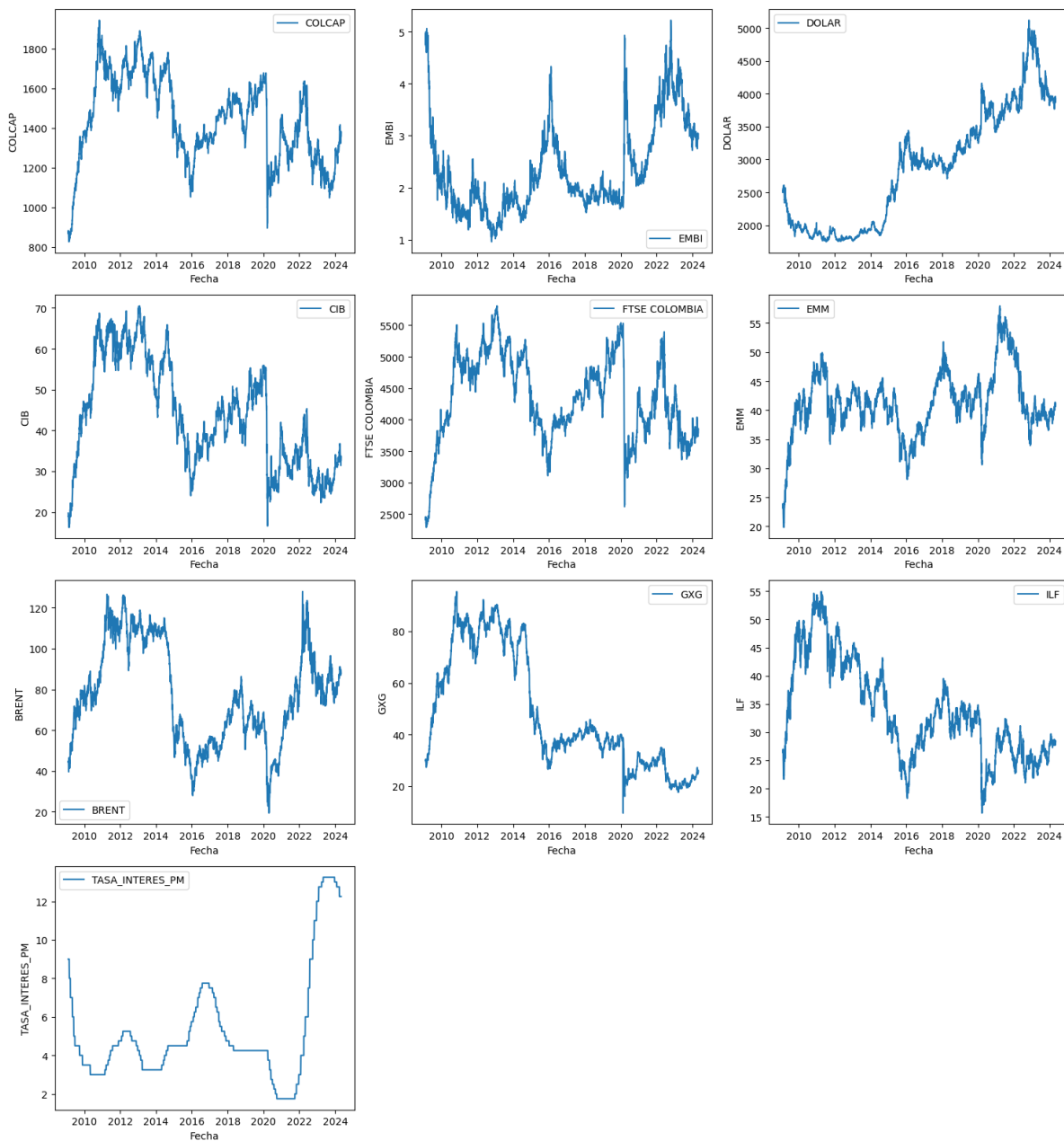


Figura 2. Gráfico de línea de las series temporales. Fuente: Elaboración propia

Estos gráficos son útiles para identificar tendencias o estacionalidades en los datos, además de proporcionar una representación visual que facilita la evaluación de si las series se comportan como una caminata aleatoria o como ruido blanco.

Al observar las imágenes, es importante considerar lo siguiente:

- **Eje X (Fecha por Año):** en todos los gráficos, el eje X representa el tiempo, distribuido por años desde 2010 hasta 2024. Esto permite identificar cambios y patrones a lo largo del tiempo.
- **Eje Y (Valores de las Series):** el eje Y varía según la serie, mostrando valores específicos para cada indicador económico o financiero.

En estos gráficos, se observa que la mayoría de las series, como COLCAP, CIB, FTSE Colombia, GXG, EEM e ILF, presentan fluctuaciones considerables a lo largo del tiempo sin una tendencia clara ni una estacionalidad evidente, lo cual es indicativo de un comportamiento errático típico de una caminata aleatoria. Estos índices muestran picos y valles frecuentes, reflejando la naturaleza volátil de los mercados financieros.

Por otro lado, el DOLAR exhibe una tendencia al alza a largo plazo con algunas caídas temporales, lo que sugiere una caminata aleatoria, pero con una dirección general creciente. El EMBI también muestra un comportamiento errático con picos pronunciados y caídas, sin una estacionalidad clara.

El precio del petróleo BRENT y la TASA DE INTERES PM por su parte se distinguen por presentar ciertos ciclos de alza y baja.

En general, las series analizadas muestran comportamientos estocásticos, característicos de caminatas aleatorias, las cuales no son fáciles de predecir por medio de movimientos claros.

Análisis de correlación

Se calculó la matriz de correlación de Pearson entre el COLCAP y las series exógenas para identificar relaciones lineales entre ellas.

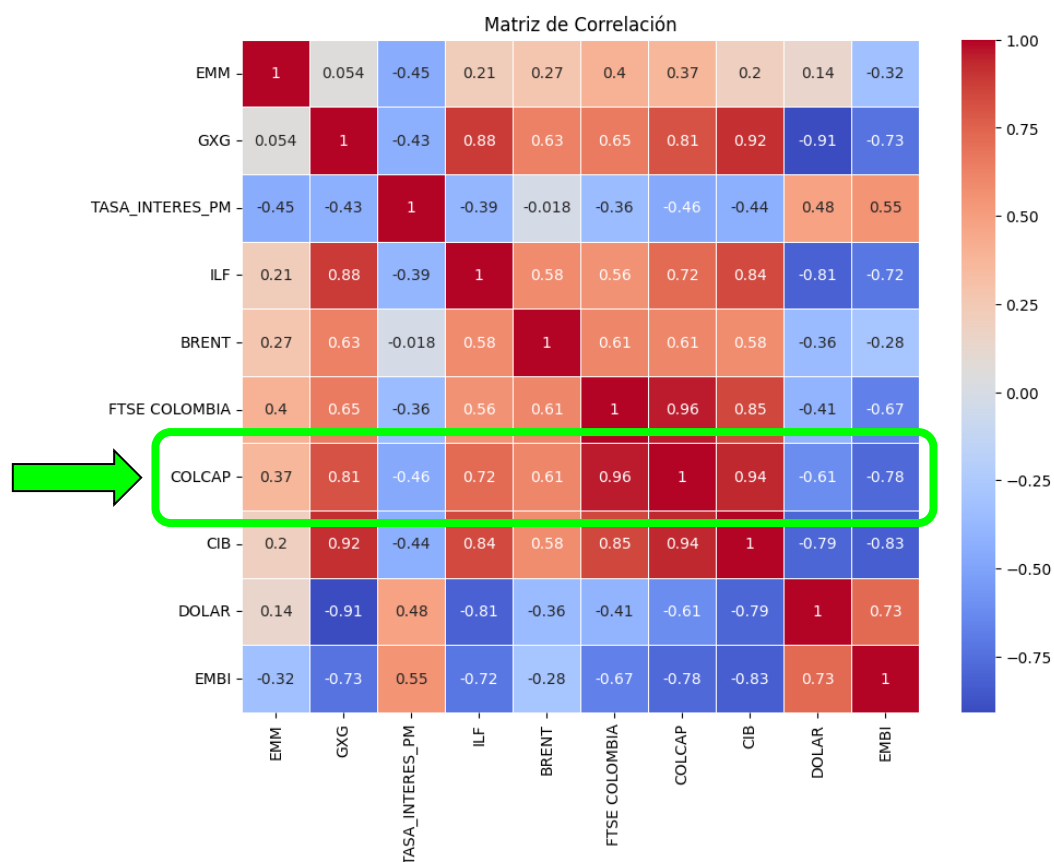


Figura 3. Análisis de correlación. Fuente: Elaboración propia.

Las correlaciones más fuertes para el COLCAP se observan con el FTSE COLOMBIA, CIB y GXG con valores de 0.96, 0.94 y 0.81, respectivamente, además de una correlación negativa de -0.78 con la variable EMBI. Estas correlaciones indican que el comportamiento del índice COLCAP está fuertemente influenciado por estos índices y variables, lo que sugiere una alta interdependencia entre los mercados financieros colombianos y los movimientos del FTSE COLOMBIA, así como de las acciones representadas por CIB y GXG.

Las demás variables también presentan correlaciones importantes, superando el 50%, a excepción de EMM y TASA_DE_INTERES_PM, cuyos resultados fueron de 0.37 y -0.46 respectivamente. Esto sugiere que, aunque estas variables tienen algún grado de relación con el COLCAP, su impacto es menos significativo en comparación con las variables antes mencionadas.

Este análisis de correlación permite corroborar la importancia de que las variables elegidas para el análisis sean relevantes, ya que una alta correlación sugiere que estas variables pueden tener un poder predictivo significativo sobre el COLCAP. Además de validar la selección de variables, este análisis nos ayuda a entender mejor las relaciones dinámicas y posibles influencias externas sobre el índice bursátil.

Por ejemplo, la fuerte correlación con el FTSE COLOMBIA puede indicar una relación directa con los mercados internacionales, mientras que la correlación negativa con el EMBI sugiere que el riesgo país también juega un papel crucial en el comportamiento del COLCAP.

Entender estas correlaciones es crucial para la construcción de modelos predictivos precisos y robustos, ya que permite identificar cuáles variables deben ser consideradas en el modelado. En resumen, el análisis de correlación no solo ayuda a validar la relevancia de las variables seleccionadas, sino que también proporciona una base sólida para la construcción de modelos predictivos eficaces y para una mejor comprensión de las fuerzas que afectan al índice COLCAP.

6.2 Análisis de estacionariedad

Una serie temporal es estacionaria si sus propiedades estadísticas, como la media y la varianza, son constantes en el tiempo. Para los modelos de series de tiempo como ARIMA y SARIMAX es fundamental contar con series estacionarias. Para evaluar si las series en estudio son estacionarias o no, se aplica la prueba de Dickey-Fuller Aumentada (ADF). Se obtuvieron los siguientes resultados:

Serie	ADF statistic	p-value	Critical Value (1%)	Critical Value (5%)	Critical Value (10%)
COLCAP	-3.2299	0.0183	-3.4319	-2.8622	-2.5671
EMBI	-3.5958	0.0058	-3.4320	-2.8622	-2.5671
DOLAR	-0.7041	0.8456	-3.4320	-2.8622	-2.5671
CIB	-2.4623	0.1249	-3.4319	-2.8622	-2.5671
FTSE COL	-3.4739	0.0086	-3.4319	-2.8622	-2.5671
EMM	-4.0839	0.0010	-3.4320	-2.8622	-2.5671
BRENT	-2.1564	0.2224	-3.4320	-2.8622	-2.5671
GXG	-1.022	0.7450	-3.4319	-2.8622	-2.5671

ILF	-2.0305	0.2733	-3.4320	-2.8622	-2.5671
<u>TASA_INTER</u>	-0.2167	0.9364	-3.4320	-2.8622	-2.5671

El análisis de estacionariedad utilizando la prueba de Dickey-Fuller Aumentada (ADF) arroja resultados mixtos para las series temporales evaluadas. Para el índice COLCAP el valor p es de 0.018, lo que es menor que el umbral de 0.05, permitiéndonos rechazar la hipótesis nula de la presencia de una raíz unitaria. Esto sugiere que la serie COLCAP es estacionaria. Similarmente, las series EMBI, FTSE Colombia y EMM presentan valores p menores que 0.05 (0.0058, 0.0086 y 0.0010 respectivamente), indicando que estas series también son estacionarias. Por otro lado, las series del DOLAR, CIB, Brent, GXG, ILF y la tasa de interés (TASA_INTERES_PM) no muestran evidencia suficiente para rechazar la hipótesis nula de no estacionariedad. Los valores p para estas series son mayores que 0.05, con valores de 0.8456, 0.1249, 0.2224, 0.7450, 0.2733 y 0.9364 respectivamente. Esto sugiere que estas series no son estacionarias y podrían requerir transformaciones adicionales, como la diferenciación, para ser adecuadamente modeladas.

Aunque las variables que sugieren estacionariedad tienen valores p acordes, la comparación entre el estadístico ADF y los valores críticos es muy cercana y, en algunos casos, el estadístico ADF no es menor que los valores críticos. Esto indica estacionariedad débil. Debido a esto y para asegurar la robustez del análisis, se procederá a diferenciar todas las series temporales. Esta transformación ayudará a garantizar que las series sean realmente estacionarias, mejorando así la precisión y la fiabilidad de los modelos predictivos que se desarrollarán posteriormente.

Transformación de las series

Como se indicó en el anterior punto, a las series se les realizó una diferenciación y luego se evaluaron nuevamente para determinar su estacionariedad. Se obtuvieron los siguientes resultados:

Serie	ADF statistic	p-value	Critical Value (1%)	Critical Value (5%)	Critical Value (10%)
COLCAP	-39.1485	0.0	-3.4319	-2.8622	-2.5671
EMBI	-15.3418	3.8216e-28	-3.4320	-2.8622	-2.5671
DOLAR	-19.9445	0.0	-3.4320	-2.8622	-2.5671
CIB	-41.4016	0.0	-3.4319	-2.8622	-2.5671
FTSE COL	-40.0412	0.0	-3.4319	-2.8622	-2.5671
EMM	-14.2441	1.5196e-26	-3.4320	-2.8622	-2.5671
BRENT	-29.1267	0.0	-3.4320	-2.8622	-2.5671
GXG	-75.8555	0.0	-3.4319	-2.8622	-2.5671
ILF	-25.2976	0.0	-3.4320	-2.8622	-2.5671
<u>TASA INTER</u>	-7.8997	4.2169e-12	-3.4320	-2.8622	-2.5671

Después de la diferenciación, todas las series transformadas mostraron resultados que sugieren estacionariedad, con valores p muy cercanos a cero y estadísticos ADF menores que los valores críticos. Estas transformaciones han sido efectivas, permitiendo un mejor análisis de las series bajo la condición de estacionariedad, con lo que se pretende que mejore la precisión y fiabilidad de los modelos predictivos.

6.3 Descomposición de las series en tendencia, estacionalidad y residuos.

Se realizó un análisis de descomposición de las series temporales para todas las series. A continuación, se presenta la gráfica de la variable estudiada COLCAP y un análisis general basado en los resultados de estas descomposiciones:

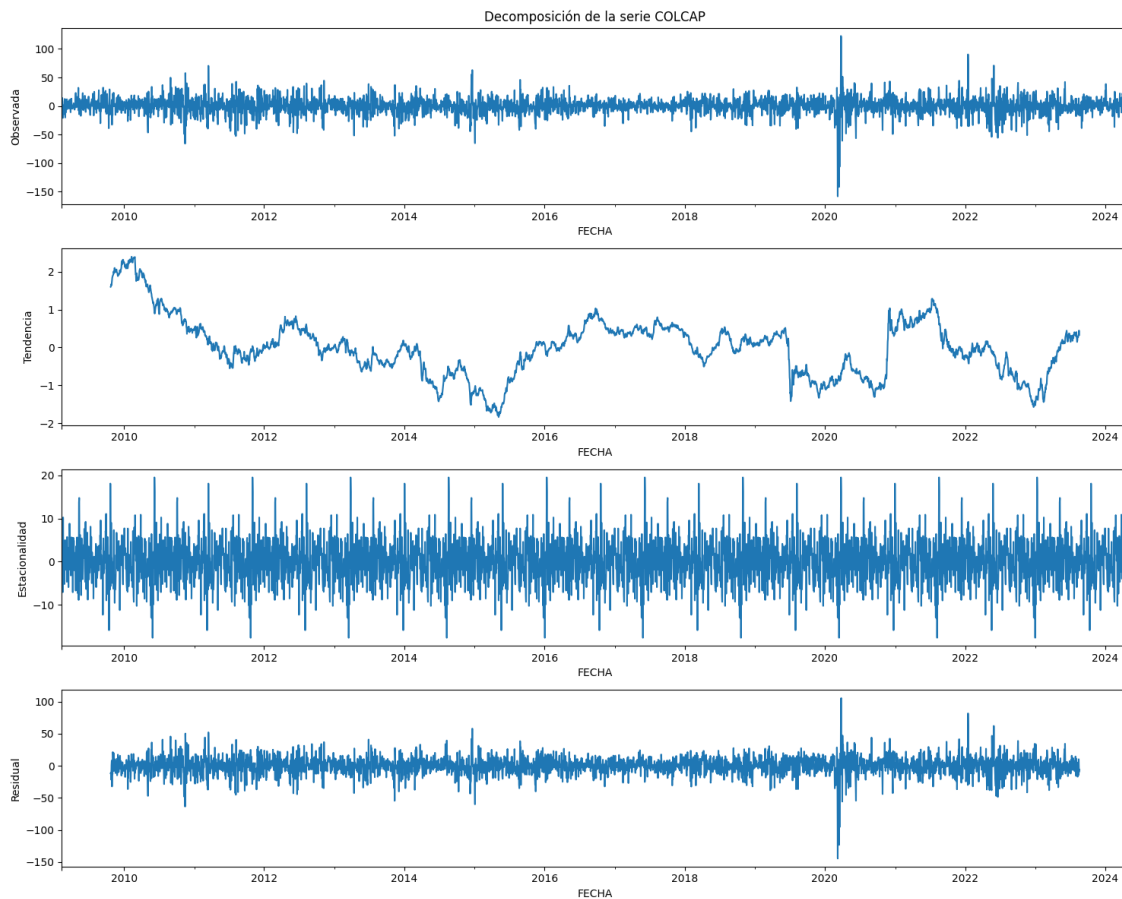


Figura 4. Descomposición de serie COLCAP diferenciada. Fuente: Elaboración propia.

Las series observadas de cada variable muestra la evolución de los datos diferenciados a lo largo del tiempo, capturando las fluctuaciones diarias y los eventos significativos que afectan a cada mercado o indicador. Al diferenciar los datos, se estabiliza la media de las series, haciendo más evidente cualquier cambio abrupto o evento significativo. Para el índice COLCAP, por ejemplo, los datos diferenciados muestran una volatilidad constante, con variaciones distribuidas uniformemente alrededor de cero, a excepción de algunos picos significativos que se pueden identificar, como el caso del 2020, que es el más notable.

La tendencia, por su parte, muestra el comportamiento tendencial de los rendimientos de cada serie temporal a lo largo de los años.

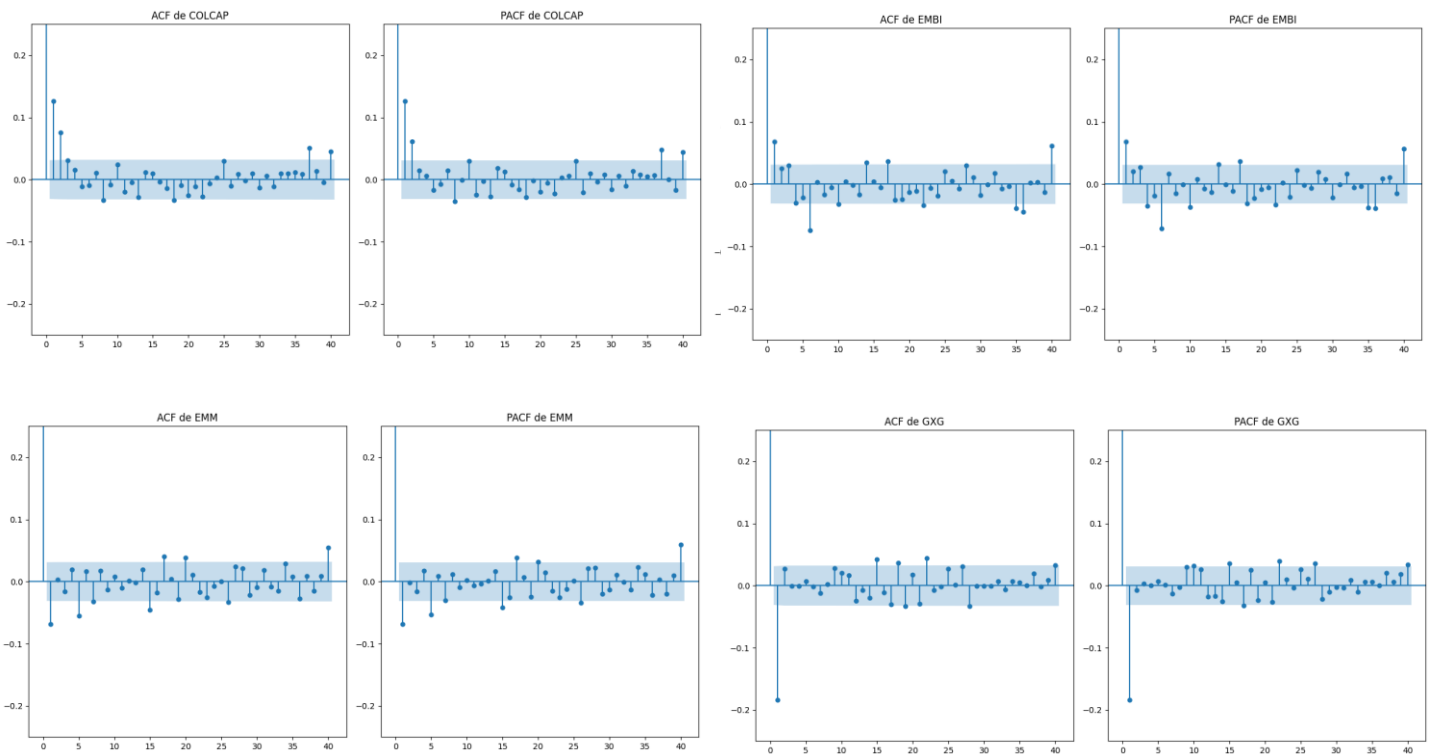
Dentro del componente estacional se pueden identificar patrones repetitivos en las series, aunque en las series de este estudio no sea clara la estacionalidad, si se pueden observar fluctuaciones periódicas bajas, que pudieran estar asociadas a factores recurrentes en el mercado financiero, como eventos económicos o cambios periódicos en las perspectivas de los inversores.

Finalmente, el análisis residual refleja una idea de las variaciones no explicadas por la tendencia o por la estacionalidad. Para la mayoría de las variables los residuos parecen estar centrados alrededor de cero con una variabilidad constante, exceptuando el movimiento que se puede evidenciar a inicios del 2020, coincidente con los eventos de la pandemia ocasionada por el COVID-19.

6.4 Análisis de Autocorrelación y Autocorrelación parcial

El análisis de autocorrelación y autocorrelación parcial permite identificar patrones de dependencia temporal. Estos patrones son esenciales para construir modelos predictivos de series de tiempo ya que informan sobre los posibles rezagos que deben ser incluidos en los modelos predictivos.

Los análisis ACF y PACF (autocorrelación y autocorrelación parcial por sus siglas en inglés) para cada una de las series se presentan en los siguientes gráficos, en donde se puede observar en el eje X los períodos de retraso y en el eje Y el nivel de autocorrelación.



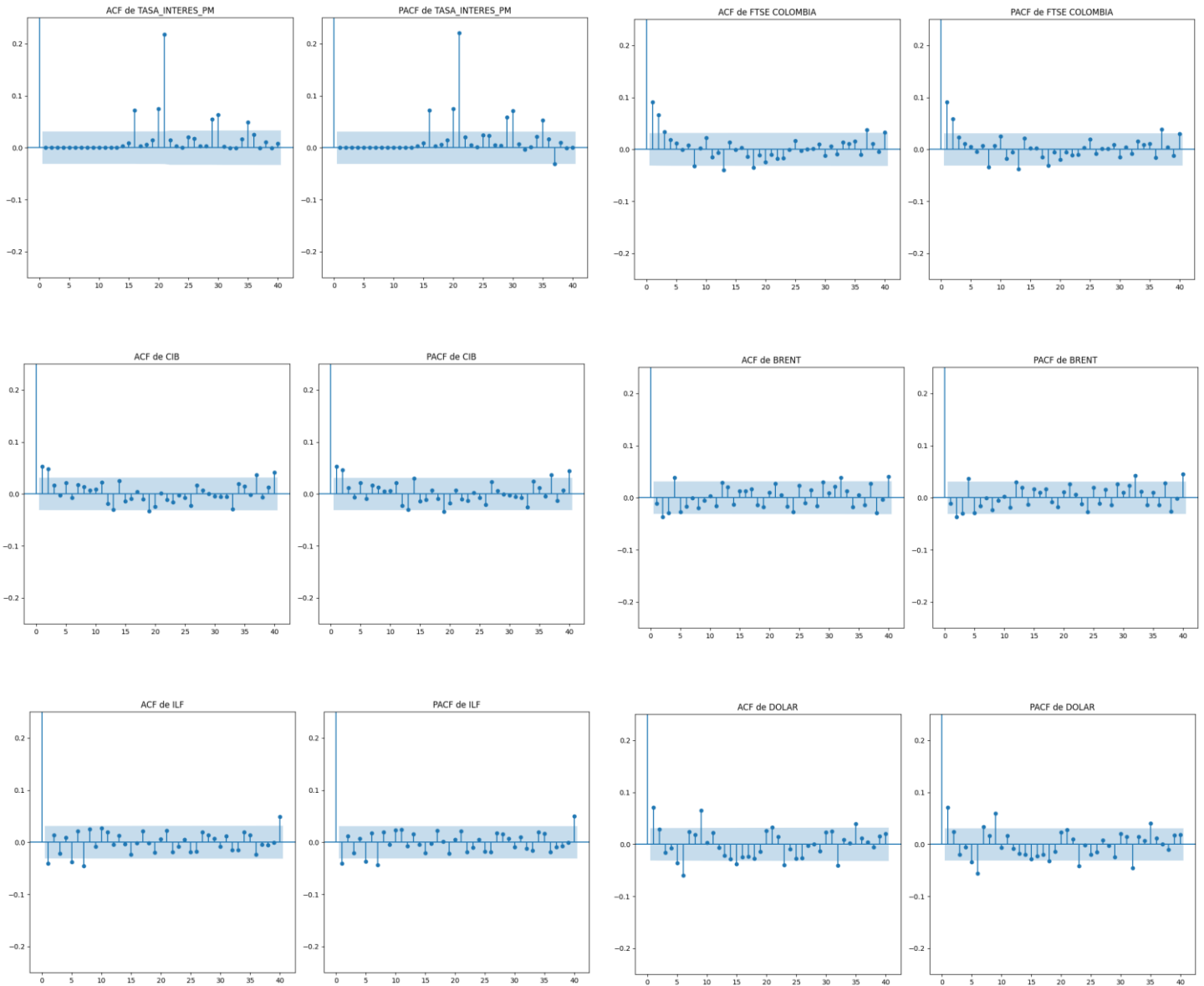


Figura 5. Análisis ACF y PACF. Fuente: Elaboración propia

Los rezagos fuera de la franja de confianza en cada una de las series, tanto en ACF como en PACF indican que dichos rezagos presentan información que debe ser tomada en cuenta en la modelación, tanto para modelos ARIMA como para modelos basados en redes neuronales.

Para las series COLCAP, CIB y FTSE COLOMBIA, se observa significancia en los coeficientes hasta el segundo rezago, lo que indica que las observaciones pasadas tienen un impacto notable en los valores actuales hasta dos períodos anteriores. En contraste, para las series DOLAR, EMBI y GXG, la significancia se limita solo al primer rezago, sugiriendo una influencia directa y más inmediata de los valores anteriores sobre el presente. Las demás variables, sin embargo, no muestran correlaciones

significativas, ya que sus coeficientes de autocorrelación se encuentran dentro de la banda de confianza, lo que implica que estas series no tengan dependencias claras en sus rezagos.

6.5 Modelación ARIMA

Análisis univariado serie COLCAP

Se realizó un análisis univariado ARIMA para la serie COLCAP diferenciada con el enfoque de auto ARIMA que minimiza el criterio de información Akaike (AIC), con el que se obtuvieron los siguientes resultados:

```
Best model: ARIMA(2,0,0)(0,0,0)[0]
Total fit time: 22.737 seconds
```

SARIMAX Results						
Dep. Variable:	y	No. Observations:	3970			
Model:	SARIMAX(2, 0, 0)	Log Likelihood	-16198.986			
Date:	Sun, 30 Jun 2024	AIC	32403.973			
Time:	22:54:52	BIC	32422.833			
Sample:	02-11-2009	HQIC	32410.661			
	- 04-30-2024					
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
ar.L1	0.1186	0.009	13.554	0.000	0.101	0.136
ar.L2	0.0611	0.008	7.975	0.000	0.046	0.076
sigma2	205.0377	2.066	99.231	0.000	200.988	209.088
Ljung-Box (L1) (Q):			0.00	Jarque-Bera (JB):		18166.60
Prob(Q):			0.95	Prob(JB):		0.00
Heteroskedasticity (H):			1.42	Skew:		-0.51
Prob(H) (two-sided):			0.00	Kurtosis:		13.43

Figura 6. Autoarima. Elaboración propia.

Este ajuste dio como resultado un ARIMA(2,0,0) indicando dos períodos en la parte autorregresiva del modelo. Este resultado es consistente con el análisis de autocorrelación y autocorrelación parcial del COLCAP, que mostró picos significativos en los dos primeros rezagos, lo cual sugiere que un AR(2) es adecuado.

Adicionalmente, se realizó un análisis ACF y PACF de los residuos del modelo por medio de la gráfica de los mismos, en donde se puede observar en el eje X los períodos de retraso y en el eje Y el nivel de autocorrelación.

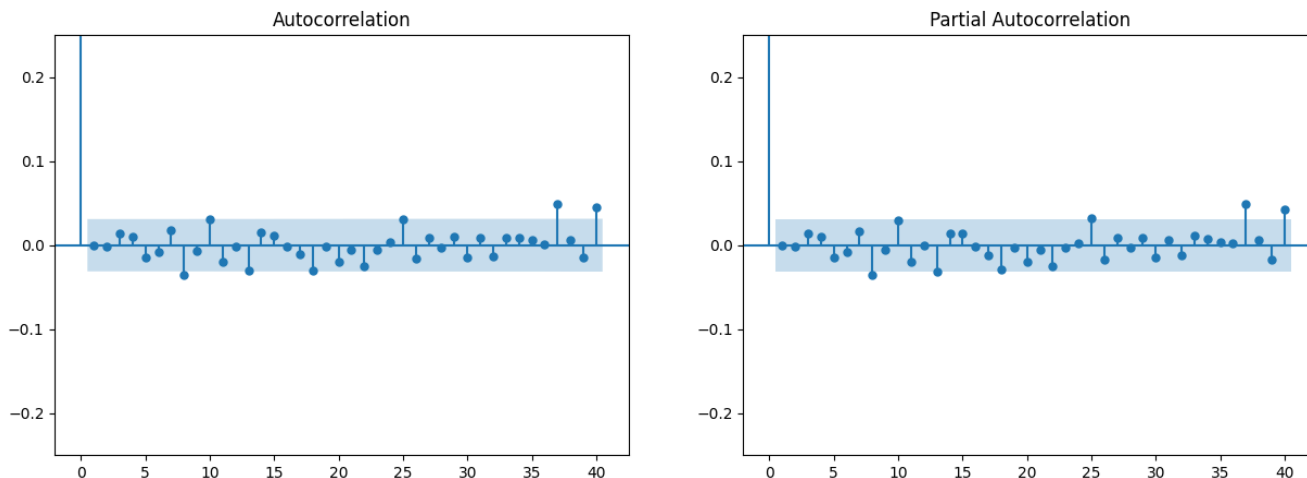


Figura 7. Autocorrelación y autocorrelación parcial de residuos modelo ARIMA.

Fuente: Elaboración propia

En la figura se puede identificar que los residuos no muestran picos significativos por fuera de la franja de confianza, sugiriendo que los residuos no presentan autocorrelación significativa. Esto apoya el resultado del análisis ARIMA de que no se necesite un componente de media móvil en el modelo.

Predicción con análisis ARIMA

Se aplicó el modelo ARIMA seleccionado en la sección anterior (ARIMA(2,0,0)) para realizar predicciones del índice COLCAP. Para este enfoque se dividieron los datos en conjuntos de entrenamiento y prueba, este último correspondiente al 20 por ciento de los periodos de la serie. Luego se ajustó el modelo a los datos de entrenamiento y se evaluó el desempeño en los datos de prueba.

Error cuadrático medio (MSE): 31611.794676889323

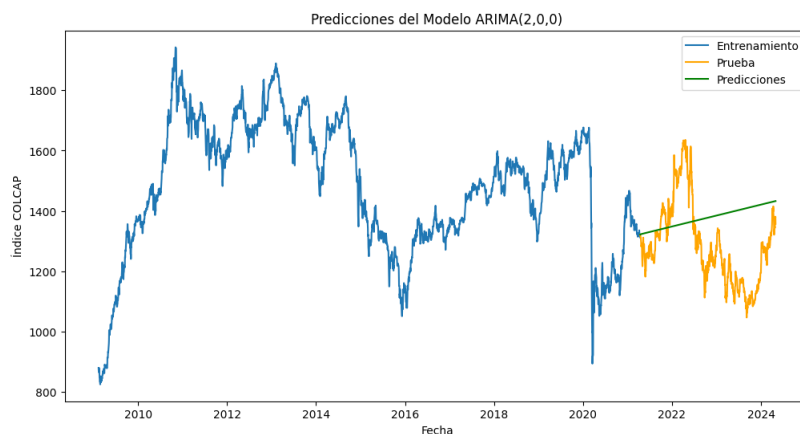


Figura 8. Resultados predicción ARIMA. Fuente: Elaboración propia.

Este análisis permitió evaluar las predicciones y graficar el resultado deficiente del modelo, pues el ajuste no es razonable y no sigue los movimientos del índice, como se puede ver gráficamente.

Predicción con análisis ARIMA y variables exógenas

Para mejorar la capacidad predictiva del modelo ARIMA previamente definido, se realizó un análisis utilizando el enfoque SARIMAX (Seasonal Autoregressive Integrated Moving Average with eXogenous variables). Este enfoque permite la incorporación de factores estacionales y variables exógenas en el modelo, lo que puede proporcionar una mayor precisión en las predicciones.

En primer lugar se partió del modelo ARIMA(2,0,0) identificado anteriormente como el mejor modelo univariado para la serie COLCAP diferenciada. Luego se consideraron varias combinaciones de variables exógenas relevantes, incluyendo EMBI, DOLAR, CIB, FTSE COLOMBIA, EMM, BRENT, GXG e ILF. El objetivo fue identificar las variables exógenas que mejoran significativamente la precisión del modelo. Posteriormente se exploraron componentes estacionales con valores de 0 y 1 para los factores P (autorregresivo estacional), D (diferenciación estacional) y Q (media móvil estacional). Además, se evaluaron periodos estacionales (S) de 5 y 7 meses para capturar posibles patrones cíclicos semanales en la serie temporal, teniendo en cuenta que los datos están en periodicidad diaria.

Se utilizó el criterio de información de Akaike (AIC) para seleccionar las mejores combinaciones de variables exógenas y componentes estacionales. El modelo con el menor valor de AIC fue considerado el mejor modelo.

Luego de realizar el análisis se obtuvo como mejor resultado un modelo SARIMAX(2,0,0)[0,0,0,5] y las variables exógenas FTSE COLOMBIA, EMM, BRENT, GXG.

Error cuadrático medio (MSE): 4229.376568163738

Gráficamente, se obtuvieron los siguientes resultados:

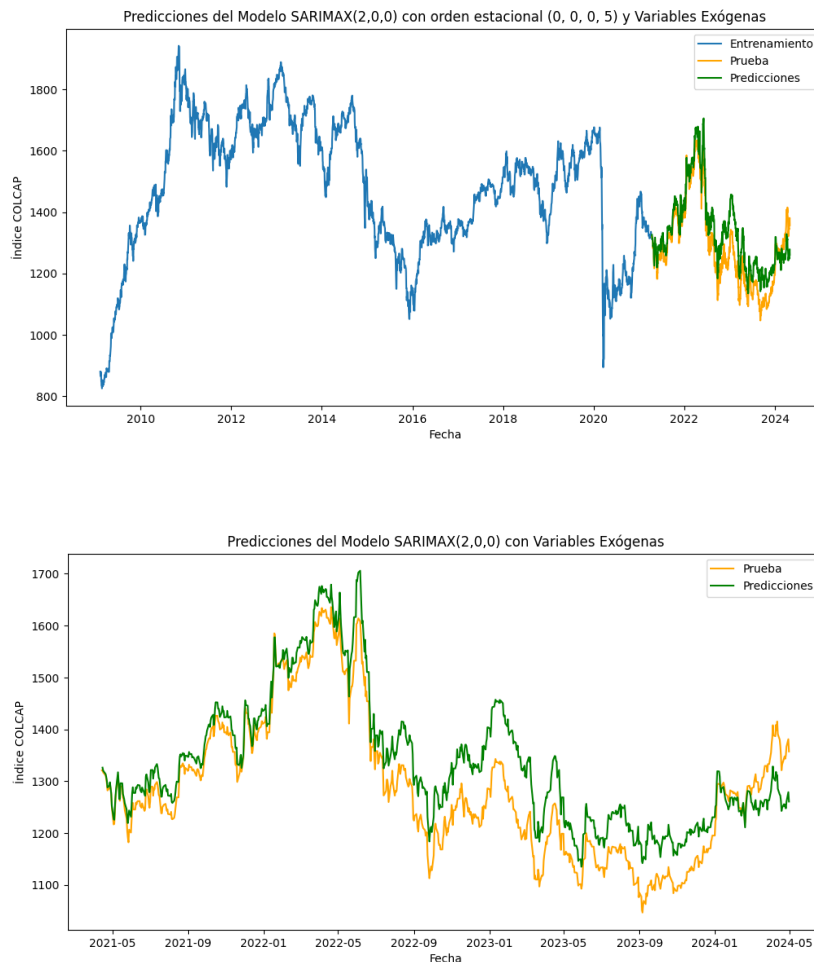


Figura 9. Resultados SARIMAX. Fuente: Elaboración propia.

El análisis SARIMAX demostró que el modelo con variables exógenas ajusta mucho mejor a la serie COLCAP que el modelo ARIMA univariado. Este modelo es muy preciso, como se evidencia gráficamente, mostrando concordancia entre las predicciones y los valores reales en el conjunto de prueba. Sin embargo, un inconveniente significativo es que, para hacer pronósticos futuros, se necesitarían los valores futuros de las variables exógenas. Por lo tanto, aunque el modelo SARIMAX es

muy efectivo para evaluar la relación histórica entre el COLCAP y las variables exógenas, su uso para predicciones a largo plazo puede estar limitado. Sin embargo, este modelo es altamente valioso para análisis retrospectivos y evaluación del impacto de factores externos en el índice COLCAP.

Modelado con redes neuronales recurrentes (RNN)

Se realiza un análisis de las series temporales utilizando Redes Neuronales Recurrentes. Específicamente se utilizó el modelo Long Short-Term Memory (LSTM) debido a su capacidad para capturar dependencias a largo plazo en series temporales.

Se utiliza la serie original de COLCAP y las series exógenas utilizadas en el punto anterior (FTSE COLOMBIA, EMM, BRENT y GXG).

Para la preparación de los datos se realiza un escalado de estos utilizando el método ‘MinMaxScaler’ para normalizarlos entre 0 y 1, facilitando la convergencia del modelo. Posteriormente, se crearon las secuencias de 60 días (aproximadamente un trimestre) para las entradas y salidas del modelo, esto con el fin de capturar las dependencias temporales. Finalmente se dividió el set de datos resultante en conjuntos de entrenamiento, validación y prueba.

Se creó un modelo LSTM que consta de una arquitectura secuencial con un total de 31.451 parámetros entrenables. La primera capa es de 50 unidades y un retorno de secuencias, seguida de una capa Dropout. La segunda capa es otra de 50 unidades, pero sin retorno de secuencias, seguida de otra capa Dropout. Finalmente, una capa densa de salida con una unidad que se utiliza para la predicción del índice.

```
Model: "sequential"
```

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 60, 50)	11200
dropout (Dropout)	(None, 60, 50)	0
lstm_1 (LSTM)	(None, 50)	20200
dropout_1 (Dropout)	(None, 50)	0
dense (Dense)	(None, 1)	51

Total params: 31451 (122.86 KB)
Trainable params: 31451 (122.86 KB)
Non-trainable params: 0 (0.00 Byte)

Figura 10. Modelo LSTM. Fuente: Elaboración propia.

El modelo LSTM entrenado mostró un buen desempeño, con un Error Porcentual Absoluto Medio (MAPE) del 1.14% en el conjunto de entrenamiento y del 1.51% en el conjunto de prueba. La pequeña diferencia entre estas métricas sugiere que el modelo generaliza bien y no presenta sobreajuste. Estos resultados indican que el modelo LSTM es efectivo para capturar las dependencias temporales y realizar predicciones precisas del índice COLCAP.

El ajuste del modelo se muestra en el siguiente gráfico:

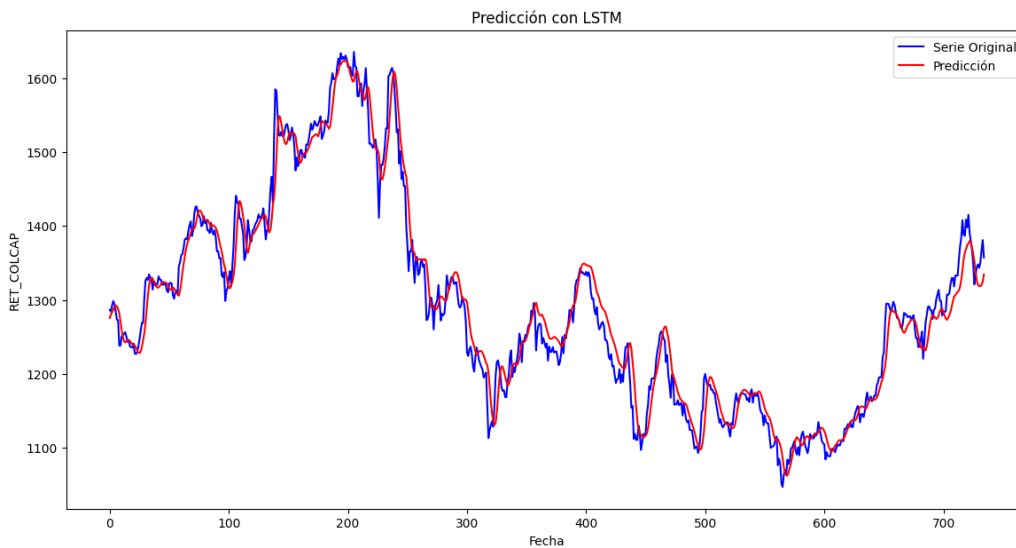


Figura 11. Resultados predicción modelo LSTM. Fuente: Elaboración propia.

6.6 Comparación de Modelos

En este proyecto se realizaron análisis utilizando los modelos ARIMA, SARIMAX y LSTM para la predicción del índice COLCAP. El objetivo fue evaluar el desempeño de cada modelo y determinar cuál proporciona las predicciones más precisas. Para la comparación de los modelos se utilizó la métrica MSE (error cuadrático medio), obteniendo los siguientes resultados:

Error cuadrático medio (MSE) ARIMA: 31611.794676889323

Error cuadrático medio (MSE) SARIMAX: 4229.376568163738

Error Cuadrático Medio (MSE) LSTM: 653.7703518069624

El modelo ARIMA fue seleccionado inicialmente debido a su simplicidad y efectividad en la modelación de series temporales univariadas. El ARIMA se enfoca en capturar las dependencias temporales dentro de la serie histórica del COLCAP sin considerar influencias externas.

Para mejorar la capacidad predictiva, se implementó el modelo SARIMAX (Seasonal ARIMA with eXogenous variables). El SARIMAX extiende el ARIMA al incluir componentes estacionales y variables exógenas, lo que permite capturar influencias externas significativas en la serie temporal. En nuestro análisis, se incorporaron las variables exógenas FTSE COLOMBIA, EMM, BRENT y GXG.

Posteriormente se implementó el modelo LSTM, el cuál destacó con un MSE significativamente menor en comparación con los modelos ARIMA y SARIMAX. Esto indica que el LSTM proporciona predicciones mucho más precisas, capturando de manera efectiva las dependencias temporales y las influencias de las variables exógenas.

Estos resultados indican que el uso de variables exógenas en los modelos es ideal para mejorar la capacidad predictiva tanto en el análisis de series de tiempo como en una arquitectura de redes neuronales recurrentes.

7. CONCLUSIONES

El análisis comparativo entre los modelos ARIMA, SARIMAX y LSTM para la predicción del índice COLCAP ha permitido identificar las fortalezas y limitaciones de cada enfoque.

El modelo ARIMA mostró limitaciones en términos de precisión. Con un Error Cuadrático Medio (MSE) de 31,611.79, el ARIMA no fue capaz de proporcionar predicciones suficientemente precisas para el índice COLCAP. Este hallazgo resalta la limitación de los modelos ARIMA cuando se aplican a una serie temporal financiera como este índice, que puede ser altamente volátil y se puede ver afectada por múltiples factores externos. Los modelos ARIMA, aunque son herramientas robustas para series temporales con patrones lineales, pueden no capturar adecuadamente la complejidad de los datos financieros. Esto implica que, para obtener predicciones más precisas del índice COLCAP, es necesario considerar modelos más avanzados.

El modelo SARIMAX, al incorporar componentes estacionales y variables exógenas (FTSE COLOMBIA, EMM, BRENT y GXG), mejoró significativamente la capacidad predictiva en comparación con el ARIMA. El MSE de 4,229.38 del SARIMAX muestra la importancia de considerar otras variables en la modelación de series temporales para índices financieros como el COLCAP.

El modelo LSTM por su parte, demostró ser el más efectivo para la predicción del índice COLCAP, con un MSE de 653.77. La arquitectura de redes neuronales recurrentes del LSTM permite una modelación flexible y precisa, superando significativamente a los modelos ARIMA y SARIMAX en términos de ajuste y predicción.

Aunque los modelos SARIMAX y LSTM pueden mostrar una predicción cercana a los valores reales del índice, se pueden ver limitados para realizar pronósticos futuros, pues se necesitaría saber el valor futuro de las variables exógenas, por lo que estos análisis permiten entender la influencia de estas variables en el índice estudiado que puedan usarse para el desarrollo de modelos predictivos más sofisticados que incorporen también el pronóstico de las variables exógenas.

8. ANEXO NOTEBOOK

Con el fin de facilitar la replicación de los resultados obtenidos en este trabajo de grado, se ha creado un repositorio en GitHub donde se encuentra el código fuente en un notebook tipo jupyter, utilizado para la ejecución de los modelos. Adicionalmente en este repositorio, se podrá acceder a los datos utilizados. El repositorio está disponible en el siguiente enlace:

https://github.com/davidsosorio/arima_sarimax_lstm

9. REFERENCIAS

Banco De La República. (2024). *BANREP*. <https://www.banrep.gov.co/es/estadisticas/>

Bhandari, H. N., Rimal, B., Pokhrel, N. R., Rimal, R., Dahal, K. R., & Khatri, R. K. C. (2022). Predicting stock market index using LSTM. *Machine Learning with Applications*, 9, 100320. <https://doi.org/10.1016/j.mlwa.2022.100320>

BLACKROCK. (2024a). *iShares Latin America 40 ETF*. <https://www.blackrock.com/cl/productos/239761/ishares-latin-america-40-etf>

BLACKROCK. (2024b). *iShares MSCI Emerging Markets ETF | EMM*. <https://www.blackrock.com/cl/productos/239637/ishares-msci-emerging-markets-etf>

Bolsa de Valores de Colombia | BVC. (2024, June 15). <https://www.bvc.com.co/msci-colcap>

Bukhari, A. H., Raja, M. A. Z., Sulaiman, M., Islam, S., Shoaib, M., & Kumam, P. (2020). Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting. *IEEE Access*, 8, 71326–71338. <https://doi.org/10.1109/ACCESS.2020.2985763>

Diaz Tagle, S., Gallego, A., Neus, C., & Sala, P. (2008). *Riesgo País en Emergentes, Mercados*.

FTSE Colombia Indices. (2024). *FTSE Russel Factsheet*. <https://research.ftserussell.com/>

Gao, P., Zhang, R., & Yang, X. (2020). The Application of Stock Index Price Prediction with Neural Network. *Mathematical and Computational Applications*, 25(3), 53. <https://doi.org/10.3390/mca25030053>

GLOBALXETFS. (2024). *MSCI Colombia ETF (GXG)*. <https://www.globalxetfs.com/funds/gxg/>

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

Mascio, D. A., Molyboga, M., & Fabozzi, F. J. (2023). The battle of the factors: Macroeconomic variables or investor sentiment? *Journal of Business and Economic Statistics*, 27(4), 417–427.

MSCI COLCAP Index Methodology. (2021). <https://www.msci.com/index-methodology>

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock Closing Price Prediction using Machine Learning Techniques. *Procedia Computer Science*, 167, 599–606. <https://doi.org/10.1016/j.procs.2020.03.326>