

Gestión de Portafolios en el Mercado Bursátil Colombiano: Comparación entre Metodologías  
Tradicionales de Optimización y Algoritmos de Machine Learning.

Rafael Eduardo Arévalo Navarro

Trabajo de Grado para Optar al Título de Economista

Director

Fernando Estrada Gallego

Universidad Industrial de Santander

Facultad de Ciencias Humanas

Escuela de Economía y Administración

Programa de Economía

Bucaramanga

2025

### **Dedicatoria**

A mis padres, quienes en todo momento me han brindado su apoyo incondicional, mostrándome que con dedicación y trabajo todo es posible.

### **Agradecimientos**

A la Universidad Industrial de Santander por brindarme las herramientas necesarias para llevar a cabo este proyecto y por contribuir a mi formación como economista.

A los maestros con quienes tuve la oportunidad de compartir a lo largo de mi carrera universitaria. En especial al profesor Fernando Estrada, quien desde tercer semestre ha confiado en mí.

A mis compañeros, con quienes tuve la oportunidad de vivir y disfrutar memorables momentos a lo largo de estos últimos cuatro años.

## Contenido

Introducción .....	11
1. Planteamiento del Problema .....	14
2. Objetivos.....	17
2.1 Objetivo General.....	17
2.2 Objetivos Específicos.....	17
3. Marco Teórico.....	18
4. Metodología .....	22
4.1 Software .....	22
4.2 Datos .....	23
4.3 Periodo de tiempo .....	25
4.4 Supuestos .....	26
4.5 Optimización Media-Varianza de Markowitz .....	26
4.5.1 Portafolio de Markowitz (Maximización Ratio de Sharpe).....	28
4.5.2 Portafolio de Mínima Varianza.....	29
4.6 Portafolio Igualitario.....	29
4.7. Modelo de Valoración de Activos Financieros (CAPM) y Alfa de Jensen .....	30
4.8 Variables .....	32
4.9 Transformación de datos.....	33
4.10 Prueba de Estacionariedad Dickey-Fuller Aumentada .....	33
4.11 Algoritmos de Machine Learning .....	34
4.11.1 Random Forest.....	38
4.11.2 Extreme Gradient Boosting (XGBoost).....	41

4.11.3 Redes Neuronales Recurrentes y LSTM.....	46
4.12 Optimización de Hiperparámetros .....	52
4.13 Validación de modelos de ML.....	52
4.14 Métricas de evaluación .....	54
4.15 Uso de las predicciones de modelos de ML.....	56
5. Limitaciones y consideraciones .....	56
6. Resultados.....	58
6.1 Aplicación de IA a la gestión de portafolios.....	58
6.1.1 Procesamiento y análisis de datos.....	59
6.1.2 Generación de predicciones .....	60
6.1.3 Optimización de asignación de capital .....	64
6.1.4 Revisiones sistemáticas de la literatura.....	66
6.2 Bolsa de Valores de Colombia.....	68
6.2.1 Servicios financieros.....	69
6.2.2 Servicios Públicos.....	70
6.2.3 Energía.....	71
6.2.4 Industria .....	72
6.2.5 Servicios y bienes de consumo .....	73
6.2.6 Materiales básicos.....	74
6.3 Índice de referencia COLCAP .....	74
6.4 Portafolios metodología de Markowitz.....	79
6.4.1 Optimización de Portafolios y Frontera Eficiente .....	84
6.4.2 Portafolio Ratio de Sharpe.....	88

6.4.3 Portafolio Mínima Varianza .....	91
6.4.3 Portafolio Ponderación Igualitaria .....	93
6.5 Portafolios Metodología Machine Learning .....	96
6.5.1 Portafolio Random Forest (RF) .....	103
6.5.2 Portafolio XGBoost (XGB) .....	105
6.5.3 Portafolio Long-Short Term Memory (LSTM) .....	108
7. Discusión.....	110
8. Conclusiones.....	117
Referencias Bibliográficas .....	122

**Lista de tablas**

Tabla 1. <i>Librerías de Python Utilizadas</i> .....	22
Tabla 2. <i>Acciones utilizadas para la construcción de portafolios</i> .....	23
Tabla 3. <i>Parámetros optimización Random Forest</i> .....	41
Tabla 4. <i>Parámetros optimización XGBoost</i> .....	46
Tabla 5. <i>Parámetros optimización LSTM</i> .....	52
Tabla 6. <i>Esquema de Validación Cruzada Temporal</i> .....	53
Tabla 7. <i>Rentabilidades anuales periodo 2019 - 2023</i> .....	80
Tabla 8. <i>Matriz de Covarianzas Anualizada (2019-2023)</i> .....	82
Tabla 9. <i>Matriz de Correlaciones (2019-2023)</i> .....	83
Tabla 10. <i>Retorno y Volatilidad Esperados</i> .....	84
Tabla 11. <i>Rentabilidades Año 2024</i> .....	87

### Lista de Figuras

Figura 1. <i>Estructura árboles de decisión</i> .....	36
Figura 2. <i>Funcionamiento neurona recurrente</i> .....	46
Figura 3. <i>Capa de neuronas recurrentes</i> .....	47
Figura 4. <i>Arquitectura celda LSTM</i> .....	49
Figura 5. <i>Comportamiento histórico índice COLCAP</i> .....	75
Figura 6. <i>Tasas de interés históricas TES 10 años</i> .....	77
Figura 7. <i>Composición sectorial histórica índice COLCAP</i> .....	78
Figura 8. <i>Frontera Eficiente</i> .....	85
Figura 9. <i>Desempeño Portafolio Ratio de Sharpe Comparado con Índice COLCAP - 2024</i> .....	89
Figura 10. <i>Ponderación Óptima Portafolio Ratio de Sharpe</i> .....	90
Figura 11. <i>Desempeño Portafolio Mínima Varianza Comparado con Índice COLCAP - 2024</i> ..	91
Figura 12. <i>Ponderación Óptima Portafolio Mínima Varianza</i> .....	93
Figura 13. <i>Desempeño Portafolio Igualitario Comparado con Índice COLCAP - 2024</i> .....	94
Figura 14. <i>Ponderación Portafolio Igualitario</i> .....	95
Figura 15. <i>Desempeño Portafolio RF comparado con Índice COLCAP - 2024</i> .....	103
Figura 16. <i>Composición Portafolio RF - 2024</i> .....	105
Figura 17. <i>Desempeño Portafolio XGB comparado con Índice COLCAP - 2024</i> .....	106
Figura 18. <i>Composición Portafolio XGB - 2024</i> .....	107
Figura 19. <i>Desempeño Portafolio LSTM Comparado con Índice COLCAP - 2024</i> .....	108
Figura 20. <i>Composición Portafolio LSTM - 2024</i> .....	109
Figura 21. <i>Comparación Ratio de Sharpe entre los diferentes Portafolios</i> .....	111
Figura 22. <i>Comparación Alfa entre los diferentes Portafolios</i> .....	112

## Resumen

**Título:** Gestión de Portafolios en el Mercado Bursátil Colombiano: Comparación entre Metodologías Tradicionales de Optimización y Algoritmos de Machine Learning\*

**Autor:** Rafael Eduardo Arévalo Navarro\*\*

**Palabras Clave:** Gestión de portafolios, Optimización de portafolios, Machine Learning, Bolsa de Valores de Colombia, COLCAP.

**Descripción:** La presente investigación evalúa la hipótesis de que los algoritmos de machine learning pueden superar el desempeño de métodos tradicionales de optimización en la construcción de portafolios de inversión en el mercado bursátil colombiano. A través de un enfoque cuantitativo comparativo, se evalúan estrategias basadas en tres algoritmos de machine learning (Random Forest, XGBoost y LSTM) comparadas con metodologías tradicionales (Markowitz, mínima varianza y ponderación igualitaria). Utilizando datos del periodo 2019-2024, fue implementada una metodología de entrenamiento y optimización con datos del periodo 2019-2023, mientras que la evaluación fuera de la muestra se realizó con datos de 2024 para 20 acciones del mercado colombiano. Los resultados revelan que el portafolio LSTM alcanza el mejor desempeño con un ratio de Sharpe de 1.5799 y un alfa de 21.40%, superando significativamente al portafolio de Markowitz (ratio de Sharpe: 0.7568, alfa: 9.96%) y al índice COLCAP. Se identifica una clara jerarquía de rendimiento relacionada con la complejidad computacional y capacidad de aprendizaje de cada algoritmo, aunque con rendimientos marginales decrecientes. Las limitaciones incluyen la dependencia temporal del análisis (año 2024), restricciones en la fuente de datos Yahoo Finance, y simplificaciones en la ejecución de transacciones. La investigación proporciona evidencia empírica de que algoritmos de machine learning pueden generar valor superior en mercados emergentes como Colombia, contribuyendo a llenar el vacío identificado en la literatura sobre aplicaciones de ML en mercados latinoamericanos, especialmente Colombia.

---

\* Trabajo de Grado

\*\* Facultad de Ciencias Humanas. Escuela de Economía y Administración. Programa de Economía.  
Director: Fernando Estrada Gallego.

**Abstract**

**Title:** Portfolio Management in Colombian Stock Market: Comparison between Traditional Optimization Methodologies and Machine Learning Algorithms\*

**Author(s):** Rafael Eduardo Arévalo Navarro\*\*

**Key Words:** Portfolio management, Portfolio optimization, Machine Learning, Colombian Stock Exchange, COLCAP.

**Description:** This research evaluates the hypothesis that machine learning algorithms can outperform traditional optimization methods in portfolio construction within the Colombian stock market. Through a comparative quantitative approach, strategies based on three machine learning algorithms (Random Forest, XGBoost and LSTM) are evaluated against traditional optimization methodologies (Markowitz, minimum variance and equal weighting). Using data from the 2019-2024 period, a methodology was implemented with training and optimization data from 2019-2023, and out-of-sample validation with data from 2024 for 20 Colombian market stocks. Results reveal that the LSTM portfolio achieves the best performance with a Sharpe ratio of 1.5799 and an alpha of 21.40%, significantly outperforming the Markowitz portfolio (Sharpe ratio: 0.7568, alpha: 9.96%) and the COLCAP index. A clear performance hierarchy correlated with computational complexity and learning capabilities is identified, albeit with diminishing marginal returns. Limitations include temporal dependency of the analysis (year 2024), constraints from Yahoo Finance data source and simplifications in transaction execution. The research provides empirical evidence that machine learning algorithms can generate superior value in emerging markets, contributing to filling the gap identified in the literature on ML applications in Latin American markets, especially Colombia.

---

\* Degree Work

\*\*Faculty of Human Sciences. Economics & Management School. Economics Program. Director: Fernando Estrada Gallego.

## Introducción

La gestión de portafolios de inversión ha presentado importantes avances desde el surgimiento de la teoría moderna de portafolios (MPT) con la publicación del reconocido trabajo de Harry Markowitz (1952) quien propuso la optimización media-varianza (MVO) como la mejor forma para construir estratégicamente portafolios de inversión diversificados. Este enfoque, complementado con aquel propuesto por el modelo CAPM (Treynor, 1962; Sharpe, 1964; Lintner, 1965), el modelo APT (Ross, 1976), y las importantes contribuciones de Fama y French (1993, 2015), han ejercido durante décadas una influencia predominante en las finanzas, y particularmente en la gestión de portafolios.

En décadas recientes el mundo ha experimentado importantes desarrollos en términos tecnológicos y de datos que han permitido una rápida mejora y mayor aplicación de los modelos de ML. Al respecto conviene mencionar el planteamiento de Petzel (2021), quien argumenta que los inversores disponen principalmente de tres mecanismos para mejorar su rentabilidad: Obtener mejor información que los demás, procesar la información mejor que los demás y contar con menores costos de operación. Particularmente, el uso de IA y ML busca brindar nuevas herramientas analíticas que cuentan con la capacidad de encontrar patrones y relaciones complejas entre numerosas variables con los cuales los gestores de portafolios y demás agentes económicos puedan tomar mejores decisiones financieras (López de Prado, 2018; Isichenko, 2021).

En términos prácticos, esto puede llevar a una revolución en el campo financiero al poner a disposición nuevas herramientas analíticas para llevar a cabo una amplia variedad de labores, permitiendo incluso el uso de nuevas fuentes de datos que históricamente eran consideradas irrelevantes o muy difíciles de aplicar en la solución de problemas financieros (Simonian et al.,

2018; Lopez de Prado, 2019), sin dejar de lado la capacidad para abordar problemas complejos al brindar una mayor flexibilidad respecto a los supuestos estadísticos tradicionales (Nosratabadi et al., 2020; Bartram et al., 2021; Kumari, 2024). Lo anterior abre la puerta a su aplicación en problemas relacionados con la gestión de portafolios, en especial a intentar mejorar las predicciones de retornos y la forma en que se distribuye el capital.

En respuesta a la hipótesis de los mercados eficientes (EMH) planteada por Eugene Fama (1970), Bernstein (1999) considera que si bien es cierto que mercados como el estadounidense tiendan a ser eficientes, la dinámica bursátil que tiene lugar en este mercado puede que no sea necesariamente la misma de otros mercados, creando así posibles ineficiencias de mercado que puedan ser explotadas mediante la gestión activa. Por otra parte, en su hipótesis de los mercados adaptativos (AMH), Lo (2004) considera que la eficiencia de los mercados no una condición estática, sino que al contrario esta responde a la dinámica presente entre los agentes que en el interactúan, dando así la posibilidad de que existan ciertas ineficiencias. En este contexto, resulta tentador evaluar la capacidad de los modelos de ML para encontrar y aprovechar posibles ineficiencias o patrones en el mercado local que potencialmente lleven a rentabilidades superiores.

Particularmente, en el caso de Colombia existen muy pocas investigaciones que traten de abordar la capacidad de la IA y modelos de ML para construir portafolios de inversión con acciones listadas en el mercado local. Principalmente es posible mencionar el trabajo de Osorio y Rico (2024), quienes encuentran que el algoritmo Random Forest brinda los mejores resultados, logrando construir un portafolios con un desempeño ligeramente superior al índice de referencia COLCAP.

Dicho lo anterior, la presente investigación busca contribuir a generar mayor evidencia al comparar el desempeño que pueden tener los portafolios construidos mediante la optimización

media-varianza propuesta por Markowitz (1952) respecto a aquel de diferentes portafolios contruidos haciendo uso de algoritmos de ML, específicamente Random Forest, XGBoost y Long-Short Term Memory. Asi mismo, son presentadas ciertas consideraciones referentes al mercado bursátil colombiano que pueden limitar la aplicación y los resultados obtenidos por estrategias de inversión basadas en ML en el contexto local.

## 1. Planteamiento del Problema

A pesar de los avances teóricos del machine learning aplicado a las finanzas y la evidencia internacional sobre su efectividad, persiste una brecha en la comprensión de como estos algoritmos se desempeñan en mercados emergentes con características particulares. El mercado de valores de Colombia presenta condiciones particulares que plantean interrogantes sobre la aplicabilidad de las metodologías utilizadas en mercados desarrollados.

El índice COLCAP ha exhibido un comportamiento histórico bastante deficiente, generando una rentabilidad anual promedio de tan solo 2.16% desde su creación en 2008, mientras que los títulos de tesorería (TES) a 10 años han proporcionado un promedio de 8.65% de rentabilidad durante el mismo periodo. La presencia de esta prima de riesgo negativa contradice los postulados fundamentales de la teoría financiera moderna y sugiere la presencia de ineficiencias en el mercado que pueden posiblemente ser explotadas mediante una estrategia de inversión activa que seleccione aquellas acciones con los mejores rendimientos esperados.

En este contexto, surge la necesidad de evaluar empíricamente si los algoritmos de machine learning pueden efectivamente superar a las metodologías tradicionales de optimización en un mercado caracterizado por baja liquidez, alta concentración sectorial y rendimientos históricamente bajos. La escasa evidencia local, limitada principalmente al trabajo desarrollado por Osorio y Rico (2024), quienes encontraron mejoras marginales usando Random Forest, indican la necesidad de un análisis más detallado que evalúe múltiples algoritmos de machine learning.

Adicionalmente, la literatura no aborda de manera detallada la relación existente entre la complejidad computacional y las mejoras de desempeño otorgadas por los diferentes algoritmos en el contexto específico de mercados emergentes. Mientras que estudios como los desarrollados por Fischer y Krauss (2018), Zhang et al. (2021) y Masuda (2024) han demostrado la superioridad

de algoritmos complejos en mercados desarrollados, no existe evidencia empírica sobre si esta relación se mantiene en mercados como el colombiano.

Al respecto de esto último, revisiones sistemáticas de literatura como las desarrolladas por Nosratabadi et al. (2020), Nazareth y Reddy (2023), Kumari (2024), Gao et al. (2024) y Sutiene et al. (2024), revelan una considerable concentración en estudiar la aplicación de técnicas de ML en mercados desarrollados, mientras que los mercados emergentes reciben una menor atención a pesar de las oportunidades que pueden llegar a ofrecer. Esto no corresponde únicamente a razones académicas, sino que puede llegar a tener implicaciones prácticas para los gestores de portafolios locales que se enfrentan día tras día a un entorno volátil donde las metodologías tradicionales pueden enfrentar limitaciones.

En este contexto surge la pregunta central de investigación: ¿En qué medida los algoritmos de machine learning (Random Forest, XGBoost y LSTM) pueden superar las metodologías tradicionales de optimización de portafolios (Markowitz, mínima varianza y ponderación igualitaria) en términos de rentabilidad absoluta y rentabilidad ajustada al riesgo en el mercado colombiano durante el año 2024? Junto a esta pregunta central es posible plantear las siguientes preguntas derivadas:

- ¿Cuál es el estado actual de la literatura académica sobre aplicaciones de machine learning en la gestión de portafolios, prestando especial atención a mercados emergentes como el colombiano?
- ¿Qué nivel de precisión logran los algoritmos Random Forest, XGBoost y LSTM en la predicción de los retornos diarios de 20 acciones listadas en la Bolsa de Valores de Colombia medidos a través del RMSE, MSE y MAE?

- ¿Cómo se comparan los portafolios construidos mediante algoritmos de machine learning versus los portafolios construidos con metodologías tradicionales al usar el ratio de Sharpe y alfa de Jensen como métricas principales de evaluación?
- ¿Existe relación alguna entre la complejidad computacional de los algoritmos y las mejoras en la rentabilidad de los portafolios?
- ¿Cuáles son las limitaciones prácticas que afectan la implementación exitosa de estrategias de machine learning en la gestión de portafolios en el contexto del mercado colombiano?

Dicho lo anterior, la hipótesis de la presente investigación establece que los algoritmos de machine learning generarán portafolios con rentabilidades superiores a aquellos portafolios construidos mediante metodologías tradicionales, siendo aquellos algoritmos con mayores capacidades los que ofrezcan un mejor desempeño.

El alcance metodológico se enfoca en considerar las 20 acciones ordinarias listadas en la Bolsa de Valores de Colombia que cuenten con información desde el año 2019, las cuales serán utilizadas para conformar los diferentes portafolios. Los datos de 2019-2023 serán utilizados para llevar a cabo el entrenamiento y optimización de los modelos, mientras que los datos de 2024 serán utilizados para la evaluación de resultados. Esta delimitación temporal permite una adecuada evaluación de desempeño en datos por fuera de la muestra, a la vez que garantiza la relevancia práctica al incluir la mayor parte de la capitalización bursátil del mercado colombiano. Las limitaciones del estudio incluyen no considerar los costos de transacción, la liquidez de cada uno de los activos y el uso exclusivo de datos de precios y volúmenes, estableciendo así condiciones controladas que permiten generar conclusiones sobre la efectividad de los algoritmos implementados a la vez que proporciona una base para futuras investigaciones más detalladas.

## 2. Objetivos

### 2.1 Objetivo General

Evaluar comparativamente el desempeño de portafolios de inversión contruidos mediante algoritmos de machine learning versus metodologías tradicionales de optimización en el mercado accionario colombiano, utilizando datos correspondientes al periodo 2024 y métricas financieras para determinar su efectividad.

### 2.2 Objetivos Específicos

Revisar la literatura académica reciente sobre la aplicación de algoritmos de machine learning en la gestión de portafolios y predicción de retornos en mercados financieros.

Implementar algoritmos de Random Forest, XGBoost y Long-Short Term Memory para la predicción de retornos diarios de las acciones listadas en la Bolsa de Valores de Colombia, utilizando datos históricos de precios y volúmenes.

Construir portafolios de inversión basados en las predicciones de los algoritmos de machine learning y compararlos con portafolios tradicionales (Markowitz, mínima varianza y ponderación igualitaria) mediante técnicas de backtesting con datos del año 2024.

Analizar el desempeño relativo de las estrategias implementadas utilizando como métricas financieras el ratio de Sharpe y alfa de Jensen para determinar su capacidad de generar valor adicional.

Identificar las principales limitaciones metodológicas y consideraciones prácticas para la implementación de modelos de machine learning en la gestión de portafolios en el mercado bursátil colombiano.

### 3. Marco Teórico

En la actualidad, son múltiples los autores que consideran que el campo de las finanzas está experimentando una revolución dados los avances recientes de nuevas tecnologías como lo es la IA (López de Prado, 2018; Simonian et al., 2018; Bartram et al., 2020, 2021). Esta transformación impulsada por la tecnología está llevando a cambios importantes en la forma en que se lleva a cabo la construcción y gestión de portafolios de inversión, no solo al complementar los enfoques tradicionales sino incluso al superarlos.

La teoría moderna de portafolios (MPT) remonta sus orígenes a la publicación del artículo “*Portfolio Selection*” por parte de Harry Markowitz (1952), quien propone la optimización media-varianza (MVO) como la mejor forma para diversificar estratégicamente portafolios de inversión. Esta tiene en cuenta la correlación existente entre los activos y con base en ella determinar la combinación que permita maximizar el retorno esperado dado cierto nivel de riesgo, o minimizar el nivel de riesgo dado cierto retorno esperado.

Años más tarde, los trabajos desarrollados por Sharpe (1964), Lintner (1965) y Treynor (1962) constituirían el modelo “*Capital Asset Pricing Model*” (CAPM), el cual propuso a  $\beta$  (la volatilidad de una acción o portafolio respecto al mercado) como un factor que podría explicar la rentabilidad de los portafolios de inversión. Siguiendo esta misma línea, Ross (1976) propuso el “*Arbitrage Pricing Theory*” (APT), en el cual el retorno de los activos no está determinado por un único factor sino por varios de ellos, siendo utilizados diferentes factores macroeconómicos. Al respecto de lo anterior, Fama & French (1993) plantean su modelo de tres factores, siendo estos beta, SMB y HML. Finalmente, Fama y French (2015) verían la necesidad de realizar ajustes a su modelo original, planteando así el modelo de cinco factores, el cual incorpora, además de los tres

factores ya mencionados, los factores de RMW y CMA. Estos autores han constituido el marco analítico predominante en las finanzas durante las últimas décadas.

Por su parte, las finanzas computacionales, también conocidas como finanzas cuantitativas, es la rama encargada de estudiar los problemas donde convergen las matemáticas, la estadística y los métodos computacionales para resolver complejos problemas financieros, en este caso aquellos relacionados con la recopilación y análisis de datos, así como el diseño e implementación de estrategias de inversión acordes a los portafolios diseñados. Actualmente, es posible apreciar las implicaciones y recientes avances de las finanzas computacionales en el uso de técnicas basadas en IA, específicamente en ML (Heaton et al., 2016; López de Prado, 2018; Nosratabadi et al., 2020; Bartram et al., 2020; Isichenko, 2021; Bartram et al., 2021; Géron, 2022; Nazareth & Reddy, 2023; Kumari, 2024), las cuales afirman tener la capacidad de superar el desempeño de técnicas tradicionales.

Conviene tener en cuenta el debate referente a la eficiencia del mercado. Mientras que Eugene Fama (1970) en su hipótesis de los mercados eficientes (EMH) sostiene que los precios ya reflejan toda la información disponible, Bernstein (1999) considera que tal hipótesis puede ser cierta para mercados como el estadounidense, más no para mercados menos desarrollados. Años después, Lo (2004) plantea la hipótesis de los mercados adaptativos (AMH) resaltando que la eficiencia no es algo dado, sino que responde a las dinámicas presentes en el mercado, generando la posibilidad de que existan periodos con cierto grado de ineficiencia que ofrezcan oportunidades de inversión. La perspectiva anterior ofrece un marco que justifica la aplicación de técnicas de ML para mejorar la detección de oportunidades de inversión que lleven a portafolios con un mejor desempeño.

A pesar de la importancia fundamental de la teoría moderna de portafolios en el desarrollo de las finanzas modernas, la optimización media-varianza presenta limitaciones significativas que han sido ampliamente documentadas en la literatura académica. La elevada sensibilidad a errores de estimación constituye una de las principales debilidades del modelo de Markowitz, donde pequeñas variaciones en los retornos esperados o en la matriz de covarianzas pueden resultar en composiciones de portafolio completamente diferentes (López de Prado, 2018).

Adicionalmente, el modelo asume distribuciones normales de retornos y correlaciones estables entre activos, supuestos que frecuentemente se violan en las dinámicas reales de mercado (Simonian et al., 2018; Ahmed et al., 2022). Al respecto de estas limitaciones, el trabajo desarrollado por Michaud (1989) demostró que MVO es extremadamente sensible a errores de estimación en los parámetros, mientras que DeMiguel et al. (2009) evidenciaron que incluso las estrategias de ponderación igualitaria pueden superar consistentemente a los portafolios optimizados al ser evaluados en una muestra diferente a la que fue utilizada para su optimización.

Los fundamentos teóricos que justifican la aplicación de machine learning en la gestión de portafolios se sustentan en la capacidad superior de estos algoritmos para superar las limitaciones estructurales de los modelos lineales tradicionales. Principalmente, los algoritmos de ML tienen tanto la capacidad como la flexibilidad para identificar relaciones no lineales entre las variables, una característica que los distingue de los enfoques econométricos convencionales (Bartram et al., 2020, 2021)

Mientras dichos enfoques tradicionales asumen relaciones lineales entre los retornos y las diferentes primas de riesgo, los mercados exhiben dinámicas totalmente alejadas de la linealidad (Petzel, 2021). Los algoritmos de ML, particularmente aquellos basados en árboles de decisión como Random Forest y XGBoost, así como arquitecturas de redes neuronales como LSTM, pueden

identificar tales relaciones complejas sin requerir especificaciones sobre la forma funcional (Géron, 2022).

Así mismo, la capacidad de procesar información multidimensional de manera simultánea representa otra ventaja teórica significativa. Como destacan Ahmed et al. (2022) y Kumari (2024), los algoritmos de ML pueden incorporar eficientemente datos alternativos y no estructurados, como lo son la información textual, sentimiento de mercado y otras variables económicas, que tradicionalmente han sido difíciles de incorporar en los modelos de optimización clásicos. Siendo esta flexibilidad particularmente importante en un contexto actual donde la velocidad y volumen de datos disponibles supera las capacidades de procesamiento de las metodologías tradicionales.

La literatura proporciona evidencia consistente sobre la superioridad de los algoritmos de machine learning frente a los métodos tradicionales de optimización (Freitas et al., 2009; Lopez de Prado, 2016; Heaton et al., 2017; Krauss et al., 2017; Fischer & Krauss, 2018; Gumelar et al., 2020; Bartram et al., 2020, 2021; C. Zhang et al., 2021; Jang & Seong, 2023; Nazareth & Reddy, 2023; Masuda, 2024). Dichos estudios demuestran que técnicas como Random Forest, XGBoost y LSTM superan de forma recurrente los resultados de metodologías tradicionales en términos de precisión y rentabilidad en las estrategias de inversión.

## 4. Metodología

### 4.1 Software

La totalidad del análisis y código fue desarrollado en Python 3.11.13 utilizando Google Colab como plataforma de desarrollo. Esta elección estuvo basada en la amplia disponibilidad de librerías especializadas para machine learning y la accesibilidad a recursos computacionales necesarios para el entrenamiento de los algoritmos. Para garantizar la reproducibilidad de los resultados, todos los recursos computacionales utilizados en la investigación están disponibles a través del siguiente enlace:

- Carpeta de Google Drive con los archivos y modelos utilizados:  
<https://drive.google.com/drive/folders/1rrhPwlrU8YTgiSPLkGHJlQqlnlIqvG9x?usp=sharing>

Allí están disponibles tanto los archivos como el código utilizado, este último recibe el nombre de “Tesis\_Portafolios\_ML” y cuya reproducibilidad requiere de los archivos presentes en la carpeta de Google Drive. A continuación son presentadas las librerías utilizadas:

**Tabla 1. Librerías de Python Utilizadas**

Librería	Versión	Propósito
pandas	2.2.2	Manipulación y análisis de datos
datetime	3.12.4	Manipulación de fechas y series de tiempo
joblib	1.5.2	Almacenamiento y carga de modelos entrenados
matplotlib	3.10.0	Visualización de gráficos y resultados
numpy	2.0.2	Operaciones matemáticas y arrays
scipy	1.16.1	Funciones de optimización y estadística
seaborn	0.13.2	Gráficos y matrices de correlación
sklearn	1.6.1	Algoritmos y métricas de ML
statsmodels	0.14.5	Análisis y pruebas estadísticas
tensorflow	2.19.0	Construcción y entrenamiento de redes LSTM
xgboost	3.0.5	Implementación del algoritmo XGBoost
yfinance	0.2.65	Descarga de datos financieros de Yahoo

Nota: Elaboración propia

## 4.2 Datos

La muestra de datos utilizada en la presente investigación corresponde a las acciones ordinarias listadas en la Bolsa de Valores de Colombia que cuentan con información disponible desde principios del año 2019. Este periodo de seis años (2019-2024), se alinea con las recomendaciones de Swensen (2009) para una educada estimación de los parámetros en la optimización de portafolios.

La muestra final corresponde a 20 acciones que cumplen los criterios de antigüedad y disponibilidad de datos, representando las empresas con cierta trayectoria y estabilidad en el mercado accionario local.

**Tabla 2. Acciones utilizadas para la construcción de portafolios**

<b>Empresa</b>	<b>Ticker</b>
Banco de Bogotá	BOGOTA.CL
Bancolombia	BCOLOMBIA.CL
Bolsa de Valores de Colombia	BVC.CL
Canacol Energy	CNEC.CL
Celsia	CELSIA.CL
Cementos Argos	CEMARGOS.CL
Constructora Concreto	CONCRET.CL
Corficolombiana	CORFICOLCF.CL
Ecopetrol	ECOPETROL.CL
Fabricato	FABRICATO.CL
Grupo Argos	GRUPOARGOS.CL
Grupo Bolívar	GRUBOLIVAR.CL
Grupo Energía Bogotá	GEB.CL
Grupo Éxito	EXITO.CL
Grupo Sura	GRUPOSURA.CL
ISA	ISA.CL
Mineros	MINEROS.CL
Nutresa	NUTRESA.CL
Promigas	PROMIGAS.CL
Terpel	TERPEL.CL

Nota: Elaboración propia con base en datos de Yahoo Finance

Los datos de precios (apertura, máximo, mínimo, cierre) y volúmenes fueron extraídos de Yahoo Finance, seleccionado como fuente principal debido a su amplia adopción en la industria financiera, su fácil accesibilidad y uso extendido en la literatura académica. Aunque Yahoo finance presenta limitaciones de granularidad (acceso únicamente a datos con frecuencia diaria) y ocasionalmente introduzca inconsistencias en los datos, proporciona la consistencia y confiabilidad necesaria para los objetivos de la presente investigación, sin dejar de lado un correcto análisis exploratorio para detectar los posibles errores. Los datos del índice COLCAP y los Títulos de Tesorería (TES) a 10 años fueron obtenidos del portal de estadísticas económicas del Banco de la República.

Durante el proceso de recopilación de datos de Yahoo Finance fueron identificados valores faltantes e inconsistencias en algunos de ellos. Para los valores faltantes estos fueron rellenados haciendo uso del promedio aritmético entre la observación anterior y posterior, preservando así la continuidad de las series de tiempo.

Las inconsistencias detectadas incluyeron cotizaciones erróneas y actividad bursátil en días festivos. Para corregir estas inconsistencias, se utilizó la información del índice COLCAP al ser considerada una fuente fiable de información directamente ofrecida por el Banco de la República para identificar los días donde efectivamente hubo actividad bursátil. Cuando fueron detectadas cotizaciones en días sin actividad en el índice COLCAP, estos datos fueron eliminados y, cuando fue necesario, se aplicó el mismo procedimiento de rellenar utilizando el valor promedio.

El anterior procedimiento garantizó la consistencia para el adecuado procesamiento de las series de tiempo por parte de los modelos implementados.

### 4.3 Periodo de tiempo

La selección del periodo 2019-2024 corresponde a consideraciones metodológicas específicas para el análisis comparativo de los diferentes portafolios. En caso de haber optado por extender el horizonte de tiempo para incluir periodos anteriores a 2019, habría sido necesario reducir el número de acciones consideradas en la muestra, limitando así el universo de activos para la construcción de los portafolios. El horizonte de cinco años, 2019-2023, utilizado tanto para el entrenamiento de los algoritmos de machine learning como para la estimación de parámetros en la optimización media-varianza de Markowitz se fundamenta en las sugerencias identificadas en la literatura para una adecuada estimación (Swensen, 2009; Petzel, 2021).

Para los algoritmos de machine learning, esta distribución temporal proporciona una división aproximada de 80-20 entre los datos de entrenamiento y evaluación, proporción que se considera óptima en la literatura para balancear el aprendizaje del modelo con su validación fuera de la muestra (Géron, 2022).

Este periodo logra incorporar la suficiente información histórica a la vez que otorga la ponderación adecuada a las condiciones experimentadas por el mercado en el corto plazo. En el caso de los modelos de machine learning, cinco años proporcionan la suficiente información para identificar patrones de movimiento. Para la optimización media-varianza, este horizonte permite calcular la matriz de covarianza y los retornos esperados con cierto nivel de confianza, evitando incluir periodos muy alejados en el tiempo o centrarse en condiciones que pueden cambiar rápidamente.

Los datos del año 2024 son utilizados exclusivamente para llevar a cabo la validación fuera de la muestra para todos los portafolios, estableciendo condiciones idénticas que permitan una

adecuada comparación entre los resultados de las técnicas de machine learning y la optimización tradicional.

#### 4.4 Supuestos

La presente investigación parte de los siguientes supuestos:

- No existen costos de transacción: Al momento de comprar o vender acciones no será cobrada ninguna comisión que pueda afectar negativamente la rentabilidad del portafolio.
- No existen costos de deslizamiento: Desde el momento en que se ejecuta una orden de venta o de compra el precio a recibir será aquel al cual cotice actualmente el mercado.
- Liquidez: El mercado contará en todo momento con la liquidez suficiente para ejecutar sin ningún problema las diferentes transacciones que requiera el portafolio.

#### 4.5 Optimización Media-Varianza de Markowitz

La optimización media-varianza (MVO) propuesta por Harry Markowitz (1952) constituye la metodología por defecto para llevar a cabo la construcción y gestión de portafolios. Tomando como fuente directa el artículo “*Portfolio Selection*”, a continuación es presentado su planteamiento matemático.

La ecuación utilizada para calcular el retorno esperado del portafolio corresponde a la siguiente:

$$E(R_p) = \sum_{i=1}^n w_i E(R_i)$$

Nota: Tomado de Markowitz (1952)

Donde:

- $E(R_p)$  corresponde al retorno esperado del portafolio
- $w_i$  es la ponderación del activo  $i$  en el portafolio

- $E(R_i)$  es el retorno esperado del activo i

Asi mismo, la varianza del portafolio está definida por la siguiente ecuación:

$$\sigma_p^2 = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_{ij}$$

Nota: Tomado de Markowitz (1952)

Donde:

- $\sigma_p^2$  es la varianza del portafolio
- $\sigma_{ij}$  corresponde a la covarianza entre los activos i y j

Modificando ligeramente la formula anterior se obtiene la desviación estándar del portafolio:

$$\sigma_p = \sqrt{\sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_{ij}}$$

Nota: Tomado de Markowitz (1952)

Por su parte, la covarianza entre dos activos ( $\sigma_{ij}$ ) se encuentra definida por la ecuación:

$$\sigma_{ij} = E \left[ (R_i - E(R_i)) (R_j - E(R_j)) \right]$$

Nota: Tomado de Markowitz (1952)

Donde:

- $R_i$  corresponde al retorno del activo i
- $E(R_i)$  corresponde al retorno esperado del activo i
- $R_j$  corresponde al retorno del activo j
- $E(R_j)$  corresponde al retorno esperado del activo j

A partir de la covarianza entre cada par de activos es posible construir la matriz de covarianzas:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

Nota: Tomado de Markowitz (1952)

Finalmente, Markowitz establece que la forma para determinar los retornos promedios, y con base en ellos determinar la covarianza entre los diferentes activos con los cuales se realizará la optimización es usar precisamente los valores diarios observados durante cierto periodo de tiempo, siendo comúnmente utilizado un periodo de 5 años (Swensen, 2009), por tal razón, el presente estudio utiliza los datos de los años 2019-2023.

#### **4.5.1 Portafolio de Markowitz (Maximización Ratio de Sharpe)**

El objetivo fundamental de la optimización media-varianza corresponde a encontrar aquella ponderación de acciones ( $w$ ) que logre maximizar el ratio de Sharpe ( $S$ ), definido por la siguiente ecuación:

$$\max_w S = \frac{E(R_p) - R_f}{\sigma_p}$$

Nota: Tomado de Sharpe (1964)

Donde:

- $E(R_p)$  corresponde al retorno esperado del portafolio
- $R_f$  corresponde a la tasa libre de riesgo
- $\sigma_p$  es la desviación estándar o riesgo del portafolio

La anterior función objetivo está sujeta a las siguientes restricciones:

- Restricción de suma de pesos:  $\sum_{i=1}^n w_i = 1$
- Restricción de no ventas en corto:  $w_i \geq 0, \quad \forall i$
- Restricción de diversificación:  $w_i \leq w_{i,\max}, \quad \forall i$ , donde  $w_{i,\max}$  representa la ponderación máxima permitida para cada acción, en este caso 20%.

#### 4.5.2 Portafolio de Mínima Varianza

La optimización media-varianza también puede ser utilizada para encontrar un objetivo diferente, siendo en este caso encontrar el portafolio que reduzca al máximo el nivel de riesgo sin realizar supuestos sobre el desempeño de cada una de las acciones. La optimización de mínima varianza está definida por la siguiente ecuación:

$$\min_w \sigma_p^2 = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_{ij}$$

Donde:

- $\sigma_{ij}$  corresponde a la covarianza entre los activos i y j
- $w_i w_j$  corresponden a las ponderaciones de los activos i y j, respectivamente.

Estando sujeta a las mismas restricciones que aquella del portafolio de Markowitz:

- Restricción de suma de pesos:  $\sum_{i=1}^n w_i = 1$
- Restricción de no ventas en corto:  $w_i \geq 0, \quad \forall i$
- Restricción de diversificación:  $w_i \leq w_{i,\max}, \quad \forall i$ , donde  $w_{i,\max}$  representa la ponderación máxima permitida para cada acción, en este caso 20%.

#### 4.6 Portafolio Igualitario

La metodología propuesta por Harry Markowitz no está exenta de errores, siendo el más común y con mayor impacto sobre las estimaciones los posibles errores que surgen al estimar la

matriz de covarianzas, siendo esta la base para llevar a cabo la diversificación del portafolio entre las diferentes acciones. Aunque existen técnicas diseñadas específicamente para reducir los errores de estimación, un procedimiento mucho más sencillo y práctico es llevar a cabo la asignación de capital de forma equitativa entre los activos que conforman el portafolio, de tal forma, se evitan dos problemas a la vez: la posibilidad de que la matriz de covarianzas presenten errores y la necesidad de realizar predicciones sobre el rendimiento esperado de las acciones, simplificando en gran medida la complejidad en la administración del portafolio (Swensen, 2009).

Aunque el portafolio igualitario no represente como tal un proceso de optimización, es conveniente presentar la ecuación utilizada para determinar la ponderación de cada acción en el portafolio:

$$w_i = \frac{1}{n}, \quad \forall i$$

Donde:

- $n$  corresponde al número de activos en el portafolio
- $w_i$  corresponde a la ponderación del activo  $i$  en el portafolio

#### **4.7. Modelo de Valoración de Activos Financieros (CAPM) y Alfa de Jensen**

El “*Capital Asset Pricing Model*”, más conocido como CAPM, fue propuesto en su conjunto por los trabajos de Treynor (1962), Sharpe (1964) y Lintner (1965). Este modelo tiene por finalidad calcular la rentabilidad esperada de cierto portafolio teniendo como referencia la volatilidad de este en relación con el mercado o índice, medida a través de la beta ( $\beta$ ), sin dejar de lado la influencia de la tasa libre de riesgo.

Antes de continuar, es conveniente presentar la fórmula utilizada para calcular la beta:

$$\beta = \frac{Cov(R_p, R_m)}{Var(R_m)}$$

Donde:

- $\beta$  corresponde a la beta del portafolio respecto al índice COLCAP
- $Cov(R_p, R_m)$  corresponde a la covarianza entre los retornos diarios del portafolio ( $R_p$ ) y del mercado ( $R_m$ )
- $Var(R_m)$  corresponde a la varianza de los retornos diarios del mercado.

A la hora de evaluar el rendimiento de cualquier fondo o portafolio de inversión, resulta conveniente separar aquel rendimiento generado gracias al movimiento del mercado y aquel rendimiento generado en “exceso” que puede ser atribuido a la estrategia de inversión implementada. Ante esto, el CAPM permite evaluar el desempeño a través de la siguiente fórmula:

$$R_p = R_f + \beta (R_m - R_f)$$

Donde:

- $R_p$  es el rendimiento del portafolio
- $R_f$  es la tasa libre de riesgo
- $\beta$  corresponde a la beta del portafolio respecto al índice COLCAP
- $R_m$  es el rendimiento del índice COLCAP

Sin embargo, la fórmula anterior permite determinar la rentabilidad teórica del portafolio, la cual puede diferir de la rentabilidad efectiva del mismo, siendo así necesario considerar el trabajo desarrollado por Michael Jensen (1967) en donde plantearía la siguiente fórmula que recibiría posteriormente el nombre de Alfa de Jensen:

$$\alpha = R_p - R_f - \beta (R_m - R_f)$$

En este sentido, cualquier portafolio construido siguiendo una estrategia de gestión activa debe generar un alfa positivo, ya que esto significa que la gestión activa logra su objetivo de ofrecer resultados superiores a los del esperados. Dicho lo anterior, alfa resulta ser una de las métricas más utilizadas al momento de evaluar la efectividad de las estrategias de inversión activa (Masuda, 2024).

De igual forma, los portafolios con un Alfa de 0 ofrecen una rentabilidad similar a la del índice de referencia tras ajustar su riesgo con base en la beta. Por otro lado, aquellos portafolios con un alfa negativo ofrecen un rendimiento peor que aquel ofrecido por el índice de referencia ajustado al riesgo.

#### 4.8 Variables

Las variables utilizadas en cada uno de los algoritmos de machine learning corresponden a las siguientes:

- Retorno logarítmico diario de la acción: Captura la rentabilidad diaria de cada acción.

Calculada a través de la siguiente fórmula:

$$R_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

- Retorno logarítmico diario del índice COLCAP: Captura la rentabilidad diaria del mercado bursátil colombiano medida a través COLCAP. Calculada a través de la siguiente fórmula:

$$R_{COLCAP,t} = \ln\left(\frac{COLCAP_t}{COLCAP_{t-1}}\right)$$

- Precio diario máximo: Precio más alto alcanzado por la acción durante la sesión de negociación. Proporciona información útil sobre la volatilidad intradía.
- Precio diario mínimo: Precio más bajo registrado por la acción durante la sesión de negociación. Junto a la variable anterior, complementa la información de la volatilidad intradía.
- Volumen de negociación: Número total de acciones transadas durante la sesión de negociación, indicando así el nivel de actividad o liquidez en un activo específico.

#### 4.9 Transformación de datos

La literatura analizada (López de Prado, 2018; Bartram et al., 2021; Géron, 2022) presenta la necesidad de realizar un preprocesamiento de datos con el fin de garantizar que estos puedan ser correctamente utilizados por los modelos de ML. Dicho esto, es aplicada la normalización de la puntuación  $Z$  para garantizar que todas las variables utilizadas se encuentren expresadas en una misma escala. Dicha normalización utiliza la siguiente fórmula:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Donde:

- $z_{ij}$  corresponde a la puntuación  $Z$  de la variable  $j$  en la fila  $i$ .
- $x_{ij}$  corresponde al valor de la variable  $j$  en la fila  $i$
- $\mu_j$  corresponde a la media de la variable  $j$
- $\sigma_j$  corresponde a la desviación estándar de la variable  $j$

#### 4.10 Prueba de Estacionariedad Dickey-Fuller Aumentada

Dado que la presente investigación utiliza series de tiempo financieras, es necesario cerciorarse de la idoneidad de estas antes de proceder con su análisis (Géron, 2022). Por tal razón,

es utilizada la prueba Dickey-Fuller Aumentada (ADF), la cual permite evaluar la presencia de raíces unitarias en series de tiempo.

La prueba ADF evalúa la hipótesis nula de que la serie presenta una raíz unitaria, respecto a la hipótesis alternativa de que la serie es estacionaria, esto a través de la siguiente fórmula:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^p \delta_i \Delta Y_{t-i} + \varepsilon_t$$

Donde:

- $\Delta Y_t$  corresponde a la primera diferencia de la serie.
- $\alpha$  es la constante.
- $\beta$  corresponde al coeficiente de tendencia temporal.
- $\gamma$  corresponde al coeficiente de la variable rezagada.
- $\delta_i$  corresponde a los coeficientes de los rezagos de las diferencias
- $\varepsilon_t$  es el término de error

Utilizando un nivel de significancia del 5%, si el p-valor es menor a 0.05 se debe rechazar la hipótesis nula y concluir que la serie es estacionaria.

#### 4.11 Algoritmos de Machine Learning

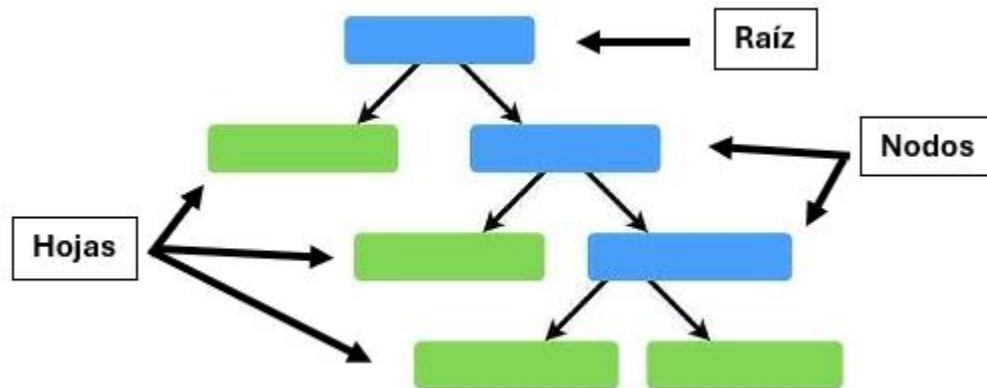
Con base en la revisión de literatura, se identificó que algoritmos como Random Forest, XGBoost y redes neuronales recurrentes (LSTM) resultan ser los más utilizados y con mejores resultados en aplicaciones financieras (Gumelar et al., 2020; C. Zhang et al., 2021; Ghosh et al., 2022; Nazareth & Reddy, 2023; Sutiene et al., 2024; Y. Zhang, 2024; Gao et al., 2024; Kumari, 2024).

Así mismo, la selección de estos algoritmos permite evaluar los resultados de diferentes enfoques de algoritmos de ML con niveles de complejidad computacional progresivos. Random Forest representa los métodos de conjunto basados en bagging con complejidad computacional relativamente baja. XGBoost incorpora técnicas de boosting secuencial con regularización avanzada y complejidad media. Por su parte, LSTM aporta la perspectiva de las redes neuronales recurrentes especializadas en secuencias temporales, representando el mayor nivel de complejidad computacional (Géron, 2022).

Esta progresión en la complejidad computacional permite evaluar si el incremento en esta última se traduce en mejoras sustanciales en el desempeño de los portafolios, proporcionando evidencia sobre la relación costo-beneficio al momento de su implementación por parte de inversores o gestores de portafolios.

Antes de hablar sobre el algoritmo de Random Forest, resulta necesario presentar el algoritmo Decision Tree (árbol de decisión). Este tipo de algoritmos pueden ser utilizados tanto para tareas de clasificación como de regresión, y gran parte de su éxito radica en el hecho de que son modelos relativamente sencillos de implementar, sin tantas de las complicaciones técnicas de algoritmos más complejos. Adicionalmente, es necesario hablar de ellos ya que son el componente principal de los algoritmos Random Forest y XGBoost. Contrario a otros modelos que suelen requerir escalar los datos, los árboles de decisión apenas requieren una preparación previa de estos (Géron, 2022), al igual que mantienen su robustez en caso de existir outliers (Lopez de Prado, 2019).

Inicialmente es necesario presentar la estructura general de los árboles de decisión:

**Figura 1. Estructura árboles de decisión**

Nota: Tomado de Josh Starmer (2022)

Partiendo de la figura 1, la parte superior de esta recibe el nombre de raíz, mientras que los demás nodos azules reciben el nombre de nodos. Por su parte, los recuadros verdes son llamados hojas, ya que son precisamente los extremos del árbol. Adicionalmente, resulta conveniente resaltar que las flechas que unen los nodos reciben el nombre de ramas, y tradicionalmente se asume que si la condición de un nodo es cierta se va hacia la izquierda, de modo que si esta es falsa se va hacia la derecha.

Específicamente los árboles de decisión funcionan a través del algoritmo conocido como “*Classification and Regression Tree*” (CART), el cual inicia dividiendo el conjunto de datos en dos partes con base en una variable  $k$  y un límite  $t_k$ , la forma de escoger cada uno de estos es mediante la búsqueda de aquel par  $(k, t_k)$  que minimice el MSE, siendo este a su vez la función objetivo. (Géron, 2022).

Matemáticamente se expresa de la siguiente forma:

$$J(k, t_k) = \frac{m_{izq}}{m} \text{MSE}_{izq} + \frac{m_{der}}{m} \text{MSE}_{der}$$

Donde:

- $J(k, t_k)$  es la función de costo para la característica o variable  $k$  en el punto de división  $t_k$
- $m_{izq}$  y  $m_{der}$  corresponden al número de observaciones en los nodos izquierdo y derecho
- $m$  es el número de observaciones en el nodo padre
- $\text{MSE}_{izq}$  y  $\text{MSE}_{der}$  son los errores cuadráticos medios en cada nodo hijo

Sin dejar de lado que:

$$\text{MSE}_{\text{nodo}} = \sum_{i \in \text{nodo}} (\widehat{y}_{\text{nodo}} - y^{(i)})^2$$

$$\widehat{y}_{\text{nodo}} = \frac{1}{m_{\text{nodo}}} \sum_{i \in \text{nodo}} y^{(i)}$$

En cuanto a sus parámetros, los árboles de decisión incluyen algunos como los siguientes:

- *Max\_depth*: Determina el número máximo de niveles que puede tener el árbol, limitando así su complejidad.
- *Min\_samples\_split*: Número mínimo de observaciones que debe tener un nodo antes de poder ser dividido.
- *Min\_samples\_leaf*: número mínimo de observaciones que debe tener una hoja para poder ser creada.

Estos últimos resultan de gran importancia a la hora de entrenar el modelo para evitar que presente problemas de sobreajuste.

Antes de continuar y finalmente hablar sobre Random Forest, es necesario una vez más definir un concepto adicional, Ensemble Learning o aprendizaje en conjunto. Comúnmente es posible considerar que los juicios emitidos por un grupo de personas tenderán a ser más precisos

que aquellos emitidos por una única persona, al fin y al cabo, la sabiduría de las masas resulta ser cierta en numerosos escenarios. Precisamente esta idea representa el sustento de los métodos de aprendizaje conjunto, los resultados de numerosos algoritmos agrupados pueden ser mejores que los de un solo algoritmo.

Existen diferentes formas de generar diversidad en los algoritmos, ya sea mediante el uso de diferentes algoritmos, o, por el contrario, entrenar un mismo algoritmo con diferentes subconjuntos de datos y variables aleatorias. Cuando este método es llevado a cabo haciendo uso del reemplazo, recibe el nombre de *Bagging*, mientras que cuando no cuenta con reemplazo recibe el nombre de *Pasting*.

Dicho lo anterior, Random Forest es un conjunto de árboles de decisión, los cuales son entrenados haciendo uso de la técnica *Bagging* (Géron, 2022). Lo novedoso de este algoritmo es que introduce cierta aleatoriedad en el proceso de generar los diferentes árboles al considerar de forma aleatoria únicamente cierta proporción de las variables para el entrenamiento, generando así diversidad que puede contribuir a mejorar los resultados. Normalmente, este valor está definido por  $\sqrt{n}$ , siendo  $n$  el número total de variables.

#### **4.11.1 Random Forest**

Matemáticamente el modelo de Random Forest aplicado a problemas de regresión luce de la siguiente forma:

Para un conjunto de  $M$  árboles, la predicción generada por el algoritmo de Random Forest es:

$$\hat{y}_i = \frac{1}{M} \sum_{m=1}^M T_m(x_i)$$

Donde:

- $\hat{y}_i$  es la predicción para la observación  $i$
- $T_m(x_i)$  es la predicción del  $m$ -ésimo árbol para la observación  $i$
- $M$  es el número total de árboles en el bosque

Muestreo con reemplazo: Cada árbol es entrenado con una muestra aleatoria del conjunto de datos:

$$\mathcal{D}_m = \text{MuestreoConReemplazo}(\mathcal{D}, n)$$

Donde:

- $\mathcal{D}_m$  corresponde al conjunto de datos para entrenar el  $m$ -ésimo árbol
- $\mathcal{D}$  corresponde al conjunto de datos original
- $n$  es el tamaño del conjunto de datos

Selección aleatoria de características: En cada nodo se lleva a cabo la selección de un subconjunto aleatorio de características:

$$\mathcal{F}_{nd} = \text{MuestreoAleatorio}(\mathcal{F}, k)$$

Donde:

- $\mathcal{F}_{nd}$  son las características consideradas para la llevar a cabo la división en un nodo
- $\mathcal{F}$  es el conjunto con la totalidad de características
- $k$  es el número de características a tener en cuenta

Criterio de división: Para aplicaciones de regresión, como la presente investigación, la división tiene por fin minimizar la suma de los errores cuadráticos para ambos nodos hijos, siendo esta a su vez la función objetivo a minimizar:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

Dicho lo anterior, la mejor división es aquella que maximiza la reducción del MSE:

$$\Delta\text{MSE} = \text{MSE}_{\text{antes}} - \left( \frac{n_{\text{izq}}}{n} \text{MSE}_{\text{izq}} + \frac{n_{\text{der}}}{n} \text{MSE}_{\text{der}} \right)$$

Importancia de variables: La importancia de cada variable es calculada con base en la disminución promedio de la impureza (varianza):

$$\text{Imp}(X_j) = \frac{1}{M} \sum_{m=1}^M \sum_{t \in \mathcal{J}_m: v(t)=j} p(t) \Delta i(t)$$

Donde:

- $\mathcal{J}_m$  es el conjunto de nodos en el árbol m
- $v(t)$  es la variable utilizada para la división en el nodo t
- $p(t)$  es la proporción de muestras que llegan al nodo t
- $\Delta i(t)$  es la reducción de impureza después de la división en el nodo t

Los hiperparámetros seleccionados controlan aspectos fundamentales del comportamiento del algoritmo:

- $n_{\text{estimators}}$ : Determina el número de árboles en el bosque. Valores más altos generalmente mejoran el desempeño, pero incrementan el tiempo de entrenamiento.

- *max\_depth*: Controla la profundidad máxima de cada árbol limitando su complejidad. Valores elevados poder causar sobreajuste mientras que valores bajos pueden limitar el aprendizaje.
- *min\_samples\_split*: Número mínimo de observaciones requeridas para dividir un nodo interno. Valores altos previenen el sobreajuste al evitar divisiones de muestras pequeñas.
- *min\_samples\_leaf*: Número mínimo de observaciones en cada hoja terminal. Valores más altos contribuyen a que el modelo no experimente problemas de sobreajuste.

Finalmente, fue definido el siguiente rango de hiperparámetros para determinar aquel que mejor se ajuste al objetivo del modelo.

**Tabla 3. Parámetros optimización Random Forest**

Algoritmo Random Forest	
Hiperparámetro	Valores
n_estimators	50 - 80 - 100
max_depth	5 - 8 - 10
min_samples_split	5 - 8 - 10
min_samples_leaf	2 - 4 - 6

Nota: Elaboración propia

#### 4.11.2 Extreme Gradient Boosting (XGBoost)

Al hablar de *Boosting* se hace referencia a la técnica de aprendizaje en conjunto que combina de forma secuencial diferentes modelos “débiles” en un solo modelo “fuerte”. La idea detrás de esto consiste en entrenar diferentes modelos secuencialmente, cada uno con el objetivo de corregir los errores o residuos generados por el modelo anterior. Uno de los modelos más utilizados corresponde a Gradient Boosting (Géron, 2022), el cual a su vez constituye la base para XGBoost.

Dicho esto, el algoritmo de XGBoost puede ser descrito por el siguiente modelo general:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in \mathcal{F}$$

Donde:

- $\hat{y}_i$  es la predicción para la observación  $i$
- $f_k$  representa el  $k$ -ésimo árbol de decisión
- $\mathcal{F}$  es el espacio de todos los árboles posibles
- $x_i$  son las características para la observación  $i$

Función objetivo: la función objetivo utilizada por XGBoost tiene en consideración dos aspectos, el error de la predicción y la complejidad del modelo:

$$\mathcal{L} = \sum_{i=1}^n l(\hat{y}_i, y_i) + \sum_{k=1}^K \Omega(f_k)$$

Donde:

- $l(\hat{y}_i, y_i)$  corresponde a la función de pérdida que evalúa la diferencia entre el valor predicho y el observado. En este caso, es utilizado el error cuadrático medio (MSE).
- $\Omega(f_k)$  corresponde al término de regularización para el  $k$ -ésimo árbol
- $n$  es el número de observaciones en el conjunto de datos

Esto permite que durante el entrenamiento el modelo optimice tanto su capacidad predictiva medida a través de la función de pérdida (MSE), así como su complejidad a través del término de regularización.

El término de regularización que utiliza el algoritmo XGBoost se define en la siguiente ecuación:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

Donde:

- $T$  es el número de hojas en el árbol
- $w_j$  es el valor de predicción para la hoja  $j$
- $\gamma$  y  $\lambda$  corresponden a parámetros de regularización

Esta resulta especialmente útil ya que controla la complejidad del modelo de dos maneras: Primero, penaliza la complejidad en la estructura de cada árbol mediante el término  $\gamma$ , y segundo, penaliza la magnitud de las predicciones mediante el término  $\lambda$ .

Para llevar a cabo la optimización de la función objetivo, XGBoost hace uso de la aproximación de segundo orden de Taylor:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^n \left[ g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

Donde:

- $g_i$  es el gradiente o derivada de primer orden
- $h_i$  es el gradiente o derivada de segundo orden
- Es la predicción en la iteración anterior.

Esto permite que el algoritmo tenga la capacidad de conocer con mayor precisión la curvatura y pendiente de la función de pérdida.

A su vez, las gradientes de primer y segundo orden que sirven de guía para llevar a cabo el proceso de optimización, al determinar tanto la dirección como la magnitud del ajuste, se ven representadas por las siguientes ecuaciones:

Gradiente de primer orden:

$$g_i = \partial_{y_i^{(t-1)}} l(y_i, \widehat{y}_i^{(t-1)})$$

Gradiente de segundo orden:

$$h_i = \partial_{y_i^{(t-1)}}^2 l(y_i, \widehat{y}_i^{(t-1)})$$

Para determinar la mejor división posible de los datos en cada uno de los árboles, XGBoost utiliza la siguiente fórmula de ganancia:

$$\text{Ganancia} = \frac{1}{2} \left[ \frac{G_I^2}{H_I + \lambda} + \frac{G_D^2}{H_D + \lambda} - \frac{(G_I + G_D)^2}{H_I + H_D + \lambda} \right] - \gamma$$

Donde:

- $G_I$  y  $G_D$  corresponden a la suma de los gradientes de primer orden en los nodos hijos izquierdo y derecho
- $H_I$  y  $H_D$  corresponden a la suma de los gradientes de segundo orden en los nodos hijos izquierdo y derecho
- $\lambda$  es el parámetro de regularización para las ponderaciones
- $\gamma$  es el parámetro de complejidad para las divisiones adicionales

En cuanto a sus implicaciones, la fórmula anterior permite no únicamente cuantificar la ganancia de implementar cierta división, sino que a la vez lo hace teniendo en consideración

el posible aumento en la complejidad del modelo, permitiendo identificar aquellas divisiones más eficientes.

Una vez determinada la estructura del árbol, el algoritmo XGBoost finalmente asigna un valor de predicción óptimo para cada hoja utilizando la siguiente fórmula:

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

Donde:

- $w_j^*$  corresponde al valor de predicción óptimo para la hoja  $j$ .
- $G_j$  es la suma de los gradientes de primer orden para la totalidad de las observaciones presentes en la hoja  $j$ .
- $H_j$  es la suma de los gradientes de segundo orden para la totalidad de las observaciones presentes en la hoja  $j$ .
- $\lambda$  es el parámetro de regularización L2

Los hiperparámetros de XGBoost controlan tanto el proceso de boosting como la regularización del algoritmo:

- *n\_estimators*: Número de árboles secuenciales en el conjunto.
- *learning\_rate*: Controla la contribución de cada árbol al modelo final. Valores menores requieren de mayor cantidad de árboles, pero pueden llevar a un mejor aprendizaje.
- *max\_depth*: Profundidad máxima de cada árbol individual. Al ser XGBoost un algoritmo secuencial es necesario definir rangos que prevengan problemas de sobreajuste.
- *min\_child\_weight*: Suma mínima de pesos requerida en un nodo hijo para permitir una división adicional. Actúa como factor de regularización al prevenir divisiones en subconjuntos con datos insuficientes.

Para concluir esta sección, fue definido el siguiente rango de hiperparámetros para determinar aquella configuración con el mejor ajuste.

**Tabla 4. Parámetros optimización XGBoost**

Algoritmo XGBoost	
Hiperparámetro	Valores
n_estimators	20 - 40 - 60
learning_rate	0,01 - 0,05 - 0,1
max_depth	2 - 4 - 6
min_child_weight	1 - 2

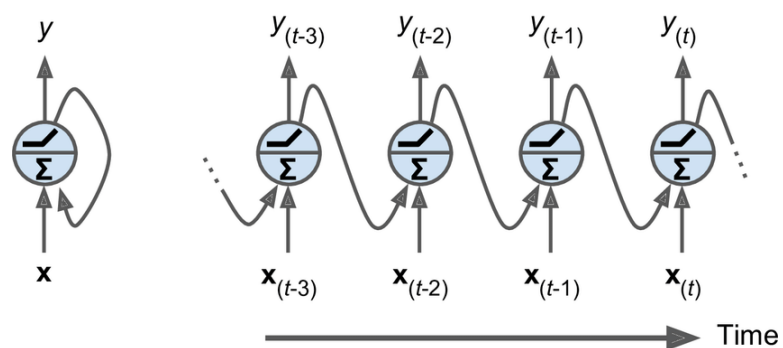
Nota: Elaboración propia

#### 4.11.3 Redes Neuronales Recurrentes y LSTM

La siguiente explicación está basada en el libro “Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow” de Aurelien Géron (2022). El lector interesado en conocer con mayor profundidad el funcionamiento de los diferentes algoritmos puede dirigirse directamente a la fuente.

La red neuronal recurrente más sencilla está compuesta por una única celda que recibe un input, procesa la información, genera un output y vuelve a recibir tal output como input en el siguiente paso. Gráficamente:

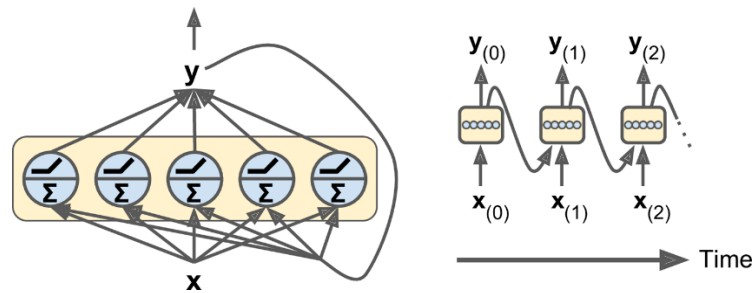
**Figura 2. Funcionamiento neurona recurrente**



Nota: Tomado de Géron (2022)

La neurona recurrente de la izquierda puede ser desenrollada a través del tiempo, lo cual mejora su representación. Básicamente, la gráfica refleja la idea de que para cada tiempo  $t$ , la red recurrente recibe no únicamente el input  $\mathbf{X}_t$ , sino que además recibe el output generado en el paso anterior ( $\mathbf{Y}_{t-1}$ ). Con base en lo anterior es posible crear capas<sup>††</sup> de neuronas recurrentes, tal como se muestra a continuación:

**Figura 3. Capa de neuronas recurrentes**



Nota: Tomado de Géron (2022)

Sin embargo, al considerar ahora diferentes neuronas cada una de ellas tiene cierta ponderación para los inputs ( $\mathbf{X}_t$ ) y otra para los outputs del paso anterior ( $\mathbf{Y}_{t-1}$ ), siendo así vectores  $\mathbf{W}_x$  y  $\mathbf{W}_y$ , de forma que el output de la capa recurrente sea calculado mediante la siguiente ecuación:

$$\widehat{y}_{(t)} = \phi(\mathbf{W}_x^T \cdot \mathbf{x}_{(t)} + \mathbf{W}_y^T \cdot \widehat{y}_{(t-1)} + b)$$

Donde:

- $\widehat{y}_{(t)}$  es el vector de outputs en el tiempo  $t$
- $\mathbf{x}_{(t)}$  es el vector de inputs en el tiempo  $t$
- $\mathbf{W}_x^T$  es la matriz de ponderaciones para los inputs
- $\mathbf{W}_y^T$  es la matriz de ponderaciones para la conexión recurrente

<sup>††</sup> Una capa es simplemente un conjunto de neuronas agrupadas en un mismo nivel.

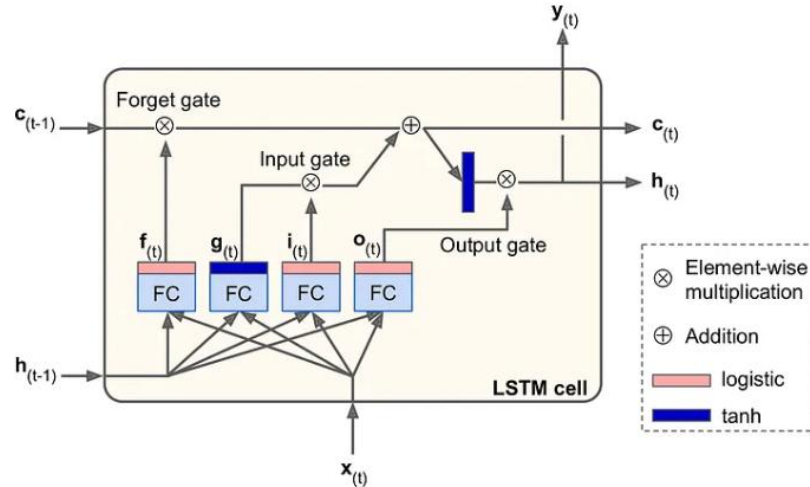
- $\widehat{y}_{(t-1)}$  es el output en el paso anterior
- $\mathbf{b}$  corresponde al vector de sesgo

Al apreciar el funcionamiento de este tipo de neurona es posible notar que tiene cierta memoria, esto al tener en cuenta los resultados obtenidos en diferentes pasos anteriores. Técnicamente, la parte de las redes neuronales que guarda información a través de diferentes pasos recibe el nombre de celda de memoria.

El estado de una celda en el tiempo  $t$  recibe el nombre de  $h(t)$ , la cual es una es una función de los inputs en el paso  $t$ , así como del paso anterior  $(t - 1)$ , de forma que:  $h(t) = f(x(t), h(t - 1))$

El proceso de entrenamiento de las redes neuronales recurrentes consiste en desdoblarlas a través del tiempo y posteriormente aplicar *backpropagation*. Técnicamente esta estrategia recibe el nombre de “*backpropagation through time*” (BPTT).

Sin embargo, aplicar RNN tradicionales a datos con secuencias de cierta magnitud puede generar problemas como gradientes inestables, la cual lleva a un entrenamiento inestable y más tardado, sin dejar de lado el hecho de que eventualmente sean olvidados los primeros inputs, limitando la memoria de la red neuronal. En respuesta a los problemas anteriores fue propuesta la celda “Long Short-Term Memory”. La arquitectura de esta celda se presenta a continuación:

**Figura 4. Arquitectura celda LSTM**

Nota: Tomado de Géron (2022)

El estado de este tipo de celda está dividido en dos vectores:  $h_{(t)}$  y  $c_{(t)}$ , los cuales pueden ser considerados como el vector de corto y largo plazo, respectivamente.

El estado de largo plazo  $c_{(t-1)}$  atraviesa la red de izquierda a derecha, pasando a través de una compuerta de olvido (forget gate), dejando atrás cierta parte de la memoria para posteriormente añadir nueva información mediante la operación de adición. El resultado de este proceso,  $c_{(t)}$ , es enviado directamente fuera de la celda. En pocas palabras, durante cada uno de los pasos cierta parte de los datos es olvidada y cierta parte es agregada. Después de la adición, es generada una copia del estado de largo plazo para someterla a la función tangente hiperbólica, cuyo resultado final es filtrado por la compuerta de salida, generando un estado de corto plazo  $h_{(t)}$ .

Al observar los demás componentes de la celda, esta recibe como input el vector  $x_t$ , así como el anterior estado de corto plazo  $h_{(t-1)}$ , los cuales son suministrados a 4 capas conectadas. La principal de todas corresponde a  $g_t$ , cuyo output es almacenado en el estado de largo plazo.

Mientras tanto, las demás capas son controladores de compuerta, ya que al utilizar la función de activación logística arrojan como resultado valores entre 0 y 1. Dado que estas luego son multiplicadas, determinan si cierran (0) o abren la compuerta (1). Específicamente:

- La compuerta de olvido ( $f_t$ ) determina aquellas partes del estado de largo plazo que deben ser olvidados.
- La compuerta de entrada ( $i_t$ ) determina aquellas partes de  $g(t)$  que deberían ser agregadas al estado de largo plazo.
- La compuerta de salida ( $o_t$ ) determina que partes del estado de largo plazo deberían ser tenidas en cuenta y mostradas en dicho paso.

La principal ventaja de esta red es que es capaz de decidir qué información de largo plazo almacenar, que información desechar y como interpretar tal información.

Dicho todo lo anterior, el funcionamiento del algoritmo LSTM se lleva a cabo mediante las siguientes ecuaciones. Para cada paso de tiempo  $t$ , con entrada  $\mathbf{x}_t$ , estado previo de la celda  $\mathbf{c}_{t-1}$ , y salida anterior  $\mathbf{h}_{t-1}$ , LSTM calcula las siguientes ecuaciones:

Compuerta de entrada (input gate):

$$i_t = \sigma(\mathbf{W}_{xi}^T \mathbf{x}_t + \mathbf{W}_{hi}^T \mathbf{h}_{t-1} + \mathbf{b}_i)$$

Compuerta de salida (output gate):

$$o_t = \sigma(\mathbf{W}_{xo}^T \mathbf{x}_t + \mathbf{W}_{ho}^T \mathbf{h}_{t-1} + \mathbf{b}_o)$$

Compuerta de olvido (forget gate):

$$f_t = \sigma(\mathbf{W}_{xf}^T \mathbf{x}_t + \mathbf{W}_{hf}^T \mathbf{h}_{t-1} + \mathbf{b}_f)$$

Contenido propuesto:

$$g_t = \tanh(\mathbf{W}_{xg}^T \mathbf{x}_t + \mathbf{W}_{hg}^T \mathbf{h}_{t-1} + \mathbf{b}_g)$$

Actualización estado de la celda:

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

Cálculo de salida:

$$y_t = h_t = o_t \odot \tanh(c_t)$$

Donde:

- $W_{xi}$ ,  $W_{xf}$ ,  $W_{xo}$  y  $W_{xg}$  representan las matrices de ponderaciones para cada una de las cuatro capas y su conexión con el vector  $x_t$ .
- $W_{hi}^T$ ,  $W_{ho}^T$ ,  $W_{hf}^T$  y  $W_{hg}^T$  representan las matrices de ponderaciones para cada una de las cuatro capas y su conexión con el estado previo  $h_{t-1}$ .
- $b_i$ ,  $b_o$ ,  $b_f$  y  $b_g$  corresponden a los vectores de sesgos para cada una de las cuatro capas.

Los hiperparámetros de LSTM controlan la arquitectura de la red neuronal recurrente y su capacidad de aprendizaje:

- *Units*: Número de neuronas LSTM en la capa oculta. Determina la capacidad de la red neuronal para capturar patrones en las series de tiempo.
- *Dropout*: Tasa de regularización que aleatoriamente “apaga” un porcentaje de las neuronas durante el entrenamiento con el fin de prevenir problemas de sobreajuste al memorizar patrones específicos.
- *Seq\_len*: longitud de la serie de tiempo que utiliza la red neuronal para hacer predicciones.

Finalmente, fue definido el siguiente rango de hiperparámetros para determinar aquel que mejor se ajuste al objetivo del modelo.

**Tabla 5. Parámetros optimización LSTM**

Algoritmo LSTM	
Hiperparámetro	Valores
units	16 - 32
dropout	0,1 - 0,2 - 0,3
seq_len	5 - 10

Nota: Elaboración propia

#### 4.12 Optimización de Hiperparámetros

Un factor importante en el entrenamiento de cualquier algoritmo de ML radica en determinar la combinación de hiperparámetros que lleve a los mejores resultados. Los hiperparámetros, tal como ya ha sido presentado, son configuraciones externas al modelo que no se aprenden durante el entrenamiento pero que controlan el comportamiento del algoritmo, influyendo directamente en sus capacidades de aprendizaje.

Dicho esto, tras definir el rango de valores para cada uno de los hiperparámetros de los diferentes algoritmos (presentados en las tablas 3, 4 y 5), fue llevada a cabo la búsqueda de la mejor combinación a través de la función GridSearchCV de la librería scikit-learn. Esta función implementa una búsqueda exhaustiva sobre el rango de hiperparámetros especificados, evaluando cada una de las combinaciones posibles y seleccionando aquella que minimiza el error, usando como métrica de evaluación el MAE.

Debido a la magnitud de la tarea a desarrollar, fue utilizado el mismo rango de hiperparámetros de cada algoritmo para todas las acciones, evitando así enfrentar mayores complicaciones computacionales y temporales que conllevaría una optimización individualizada.

#### 4.13 Validación de modelos de ML

Un factor crítico a la hora de implementar cualquier tipo de algoritmo de ML corresponde a su validación. En este sentido, al tratarse de una aplicación de series de tiempo resulta indispensable aplicar procedimientos que eviten problemas de fuga de datos (data leakage), los

cuales pueden verse reflejados en el hecho de que el modelo conozca con antelación datos que no debería conocer, mejorando artificialmente los resultados obtenidos. Por tal razón, los algoritmos Random Forest y XGBoost fueron validados haciendo uso de la validación cruzada temporal (TimeSeriesSplit), la cual preserva el orden cronológico de las observaciones.

El esquema de validación empleado es presentado a continuación:

**Tabla 6. Esquema de Validación Cruzada Temporal**

Validación Cruzada Temporal		
Fold	Entrenamiento	Validación
1	2019-2020	2021
2	2019-2021	2022
3	2019-2022	2023

Nota: Elaboración propia

Presentada la tabla anterior, una vez sean seleccionados los mejores hiperparámetros a través del esquema de validación, se procede a realizar el entrenamiento final de cada modelo para finalmente evaluar sus resultados con datos del año 2024. Tal procedimiento garantiza que en ningún momento el modelo tenga acceso a los datos del año 2024 durante su entrenamiento, evitando así que exista la fuga de datos.

Para el algoritmo LSTM se implementó un esquema de validación cruzada temporal adaptado a las características específicas de las redes neuronales recurrentes. Dado que LSTM requiere la construcción de secuencias de tiempo de longitud específica (seq\_len), fue necesario desarrollar un procedimiento de validación que preservara la integridad de las secuencias mientras evitaba la fuga de datos.

El proceso empleado fue aplicar una división temporal en tres folds progresivos: el primer fold utilizó el 50% inicial de las secuencias para entrenamiento y el 25% siguiente para validación; el segundo fold empleó el 65% inicial para entrenamiento y el 25% siguiente para validación; el tercer fold utilizó el 80% inicial para entrenamiento y el 20% restante para validación. La selección

de hiperparámetros se basó en el promedio del error absoluto medio (MAE) obtenido en los tres folds de validación.

Al igual que en los algoritmos anteriores, una vez seleccionados los mejores hiperparámetros, se llevó a cabo el entrenamiento final del modelo LSTM utilizando todas las secuencias del periodo 2019-2023 para evaluar su desempeño con secuencias construidas exclusivamente con datos de 2024.

#### 4.14 Métricas de evaluación

Teniendo en cuenta la investigación llevada a cabo por Masuda (2024), las métricas utilizadas para evaluar la precisión en las predicciones de los modelos de ML son el error cuadrático medio (MSE), raíz del error cuadrático medio (RMSE) y error absoluto medio (MAE), definidos a continuación:

Error cuadrático medio (MSE): Penaliza mayormente errores grandes, siendo especialmente útil para identificar modelos que generen predicciones muy erróneas, lo cual resulta ser muy importante en aplicaciones financieras donde un error puede afectar significativamente la rentabilidad de la estrategia.

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

Donde:

- $n$  es el número de observaciones
- $y_i$  es el valor observado
- $\hat{y}_i$  es el valor predicho

Raíz del error cuadrático medio (RMSE): Proporciona una interpretación en las mismas unidades que la variable objetivo, facilitando la evaluación práctica de la precisión de las predicciones en términos de rentabilidad esperada.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Donde:

- $n$  es el número de observaciones
- $y_i$  es el valor observado
- $\hat{y}_i$  es el valor predicho

Error absoluto medio (MAE): Ofrece una medida robusta con menor sensibilidad a los errores grandes respecto a MSE, proporcionando así una evaluación más conservadora del desempeño de los modelos.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Donde:

- $n$  es el número de observaciones
- $y_i$  es el valor observado
- $\hat{y}_i$  es el valor predicho

#### ***4.15 Uso de las predicciones de modelos de ML***

Los modelos de machine learning son entrenados con el fin de predecir el retorno de cada una de las acciones en el próximo día de negociación. Seleccionando aquellas con retornos positivos predichos para la construcción del portafolio.

Específicamente, la estrategia implementada consiste en:

1. Selección de activos: Al final de cada jornada, los modelos generan predicciones para el día siguiente, identificando las acciones con retornos positivos esperados.
2. Asignación de capital: El capital disponible se distribuye de forma igualitaria entre todas las acciones seleccionadas.
3. Ejecución: Todas las transacciones se asumen ejecutadas al precio de cierre, simplificando el cálculo y evitando las complejidades que conlleva el hecho de modelar los spreads bid-ask, así como la demás información intradía.
4. Rebalanceo diario: Una vez finalizado cada día de negociación, se incorporan los nuevos datos del mercado para actualizar los modelos y generar las predicciones para la próxima jornada de negociación.

Sin embargo, es necesario reconocer que el supuesto de ejecución al precio de cierre representa una limitación práctica importante ya que en condiciones de negociación reales las operaciones pueden ejecutarse a precios distintos a causa de factores como la liquidez.

### **5. Limitaciones y consideraciones**

A pesar de la búsqueda de información llevada a cabo es necesario mencionar ciertas limitaciones y consideraciones que pueden llegar a influir en los resultados obtenidos.

Primero, teniendo en cuenta que la Bolsa de Valores de Colombia fue establecida tan solo a principios del milenio, y que no todas las empresas listadas están presentes en la bolsa desde sus inicios, la disponibilidad y la granularidad de los datos resulta limitada para llevar a cabo análisis más detallados y con un mayor horizonte de tiempo. De igual forma, la fuente de información financiera, Yahoo Finance, debe ser revisada antes de la implementación y construcción del portafolio ya que fueron encontradas algunas inconsistencias en los datos. Esto a su vez significa que los algoritmos de ML no pueden explotar al máximo sus capacidades para trabajar con bases de datos grandes y complejas.

Segundo, los algoritmos de ML requieren de una mayor capacidad de cómputo que modelos tradicionales como la regresión lineal o logística, de allí la importancia de un hardware especializado (GPU) que pueda soportar y acelerar los diferentes procedimientos. Al respecto, el autor tuvo únicamente acceso a los servicios gratuitos de computación en la nube de Google Colab. Mayor capacidad de cómputo permitiría no solo acelerar las etapas de entrenamiento sino también trabajar con bases de datos más complejas, posiblemente llevando a mejores resultados.

En tercer lugar, la gestión de portafolios, al igual que el diseño y entrenamiento de algoritmos de ML deben ser vistos como toda una cadena de montaje (López de Prado, 2018), en los cuales cierta parte de un equipo está especializado en una labor específica, ya sea la ejecución de transacciones, construcción de modelos financieros y demás. Dicho esto, resulta extremadamente difícil para una única persona encargarse de todas estas funciones, razón por la cual fue necesario dejar de lado muchas de estas.

Finalmente, es necesario mencionar que, aunque el autor haya tenido una formación tradicional en el área de la estadística y la econometría, este constituye su primer proyecto aplicando técnicas de ML. A pesar de haber realizado un esfuerzo por entender e implementar

dichas técnicas, aún existen limitaciones de conocimiento que dejan un considerable margen de mejora a medida que dichos conocimientos evolucionen.

## **6. Resultados**

### **6.1 Aplicación de IA a la gestión de portafolios**

Los desafíos e incertidumbre que enfrenta la industria financiera actualmente han llevado a que esta deba buscar soluciones más complejas, razón por la cual ha aumentado enormemente el interés en técnicas modernas como lo es la IA (Bahrammirzaee, 2010), las cuales no solo complementan las técnicas tradicionales, sino que incluso llegan a reemplazarlas (Gao et al., 2024; Kumari, 2024).

Al respecto de ello, Marcos López de Prado (2019) presenta la idea de que si bien los mercados pueden ser eficientes en el sentido propuesto por Fama (1970) y Malkiel (2020) bajo la mira de herramientas como la econometría, es posible que estos sean lo suficientemente ineficientes al hacer uso de técnicas estadísticas modernas, tales como ML. Según López de Prado (2019), en múltiples ocasiones las bases de datos utilizadas para llevar a cabo aplicaciones financiera pueden presentar problemas como la existencia de datos no estructurados, alta dimensionalidad, dispersión de los datos e incluso relaciones complejas y no lineales entre las variables, ante los cuales las técnicas de ML ofrecen tanto la capacidad numérica así como la flexibilidad funcional.

En línea con lo anterior, Bahrammirzaee (2010) destaca como una gran ventaja de los métodos no paramétricos (Decision Trees, Random Forest y LSTM) sobre los modelos paramétricos el hecho de que los primeros no deben realizar supuestos acerca de la distribución

estadística de las variables utilizadas, lo cual permite su aplicación en una mayor variedad de problemas.

Tomando en cuenta lo planteado por Bahrammirzaee (2010), López de Prado (2018, 2019), Ahmed et al. (2022), Kumari (2024), Sutiene et al. (2024) y Gao et al. (2024), es posible mencionar los principales campos de aplicación de la IA en la gestión de portafolios:

### ***6.1.1 Procesamiento y análisis de datos***

En este apartado resulta indispensable mencionar la capacidad de la IA para procesar datos no estructurados o alternativos, como pueden serlo las imágenes, reportes, noticias, conferencias e incluso el análisis de sentimiento a través de redes sociales (Hirschberg & Manning, 2015; López de Prado, 2018; Ahmed et al., 2022; Géron, 2022; Gao et al., 2024; Kumari, 2024).

En el contexto financiero, esto significa que las bases de datos disponibles no están limitadas a información como los precios, volumen, volatilidad o variables macroeconómicas, sino que resulta posible utilizar la IA para analizar imágenes satelitales que permitan cuantificar la actividad marítima en un puerto de carga, tener en cuenta los niveles de contaminación de una zona industrial como posible indicador de la actividad productiva y mucho más.

De igual forma, también es posible utilizar modelos de procesamiento natural de lenguaje (NLP) para llevar a cabo rápidamente el análisis de informes financieros, así como el análisis e interpretación de noticias que abarquen no solo eventos económicos sino también eventos políticos con implicaciones económicas. Esto sin dejar de lado el análisis de redes sociales como X y Reddit a las cuales numerosas personas acuden para discutir sobre temas relacionados con la inversión, siendo posible extraer información sobre el posible estado o sentimiento del mercado.

### ***6.1.2 Generación de predicciones***

Este campo de aplicación es uno de los más recurrentes en la investigación académica, tal como se ve reflejado en los trabajos de Sezer et al. (2020) Nosratabadi et al. (2020), Ahmed et al. (2022), Nazareth & Reddy (2023), Gao et al. (2024) y Sutiene et al. (2024). A continuación se presenta una revisión cronológica de los principales estudios empíricos en esta área:

Freitas et al. (2009) representa uno de los primeros esfuerzos sistemáticos en aplicar redes neuronales para predicción financiera en el contexto latinoamericano. Los autores recurren a su modelo propuesto “autoregressive moving reference neural network” (AR-MRNN) para predecir el rendimiento semanal de cincuenta y dos acciones listadas en la bolsa de valores de Brasil (IBOVESPA) utilizando como periodo de tiempo octubre de 1999 a septiembre de 2007. Adicionalmente, derivan una medida de riesgo basada en los errores de predicción, a partir de la cual crean un marco para optimizar el portafolios sin dejar de lado la diversificación. En cuanto a los resultados alcanzados, el algoritmo AR-MRNN logra superar tanto al modelo MVO (optimización media-varianza) propuesto por Markowitz, así como el índice IBOVESPA, con una rentabilidad de 292 puntos porcentuales superior al modelo MVO y 78 puntos porcentuales respecto al índice IBOVESPA. Los autores concluyen que este tipo de metodologías basadas en la predicción de retornos pueden ser más adecuadas para la inversión a corto plazo.

Fisher y Krauss (2018) realizan un importante estudio al utilizar el algoritmo LSTM para predecir la dirección del movimiento de las acciones que forman parte del S&P 500, utilizando como periodo de tiempo 1992-2015. El uso de estos algoritmos les permitió obtener una rentabilidad diaria promedio de 0.46%, así como un ratio de Sharpe de 5.8, resultados que logran vencer aquellos de los modelos Random Forest, Deep Neural Network y regresión logística. Sin embargo, el uso de LSTM logra generar rentabilidades tan solo hasta el año 2009, pues a partir de

2010 dicha rentabilidad positiva se desvanece. A pesar de que la rentabilidad eventualmente termina por reducirse, un análisis de las fuentes de riesgo muestra que este portafolio presenta poca exposición a los factores de riesgo tradicionales, aportando evidencia de la capacidad del modelo LSTM para extraer patrones de movimiento en las series de tiempo financieras.

Por su parte, Ghosh et al. (2022) implementan una estrategia similar a Fischer y Krauss (2018), utilizando modelos de Random Forest y LSTM para predecir la dirección de los retornos diarios de las acciones del S&P 500 durante el periodo 1993-2018. Los resultados muestran que LSTM logra generar retornos diarios promedio de 0.64%, mientras que para Random Forest estos toman el valor de 0.54%, confirmando la superioridad de las arquitecturas secuenciales en tareas de predicción de series de tiempo.

Gumelar et al. (2020) llevan a cabo una comparación directa entre los algoritmos XGBoost y LSTM para predecir los precios de veinticinco acciones listadas en la bolsa de Indonesia. Utilizando variables como el precio de apertura, máximo, mínimo, cierre y volumen para el periodo de tiempo 2000-2019, los autores desarrollan una estrategia tanto de compra como de venta de acciones. Los resultados permiten determinar que el algoritmo LSTM supera ligeramente a XGBoost, al obtener una precisión de 99.995%, mientras que la precisión de XGBoost fue de 99%.

Zhang (2024) hace uso de un algoritmo XGBoost optimizado mediante la función GridSearchCV para predecir precios de acciones basándose en indicadores técnicos como medias móviles exponenciales (EMA), medias móviles simples (SMA), el índice de fuerza relativa (RSI) y la media móvil de convergencia/divergencia (MACD) como variables predictivas.

Los resultados demuestran que el algoritmo XGBoost logra alta precisión en la predicción de precios, siendo el indicador RSI el de mayor importancia al momento de predecir. Sin embargo,

el autor reconoce explícitamente las limitaciones del modelo, particularmente su vulnerabilidad ante eventos económicos inesperados y cambios significativos en las empresas, aspecto que pocos estudios mencionan de forma explícita en sus análisis.

Un trabajo particularmente relevante para la presente investigación es el desarrollado por Masuda (2024), quien implementa un enfoque híbrido que combina tanto modelos de machine learning con la optimización media-varianza para llevar a cabo la construcción de portafolios. Utilizando datos de las 50 empresas más grandes de Estados Unidos en el periodo 2019-2023, siendo los datos de entrenamiento aquellos de 2019-2022 y los datos de prueba aquellos del año 2023, son evaluados nueve algoritmos diferentes incluyendo modelos tradicionales (regresión lineal), modelos individuales (RF, LightGBM, LSTM) y modelos híbridos (CNN-LSTM).

Este estudio empleó 42 variables extraídas del “Center for Research in Security Prices” (CRSP), siendo necesario reducir su dimensionalidad mediante PCA y posteriormente estandarizar las variables a través de la puntuación Z. La calidad predictiva de los diferentes modelos aplicados es evaluada a través de las métricas MAE, MSE y SMAPE, mientras que los resultados de los portafolios son evaluados a través de su ratio de Sharpe y Alfa.

Los resultados revelan una jerarquía de rendimiento entre algoritmos. Al considerar el alfa, los modelos híbridos resultan ser mejores, tanto CNN-LSTM como BiLSTM-BO-LightGBM generan un alfa superior al 5%. Mientras tanto, modelos más sencillos como la regresión lineal, SVM y RNN llegan a presentar valores de alfa negativos. Al considerar el ratio de Sharpe los resultados presentan un orden similar.

Sin embargo, su investigación plantea algunas consideraciones críticas, como el hecho de que durante el año de prueba (2023) el mercado americano experimentó un retorno de 24.14%, condicional excepcional que no garantiza el buen desempeño de los algoritmos en circunstancias

de mercado diferentes. Así mismo, el hecho de considerar únicamente acciones de empresas blue-ship limita su replicabilidad en portafolios más diversificados. Finalmente, es planteado el problema de que los algoritmos más complejos requieren capacidades computacionales mayores.

El trabajo desarrollado por Masuda brinda un marco de referencia valioso para la presente investigación al generar una comparación rigurosa del desempeño que tienen los portafolios construidos con base en diferentes modelos de ML, destacando tanto las oportunidades como las limitaciones prácticas de estos enfoques en condiciones reales de mercado.

Finalmente, en el contexto local Osorio & Rico (2024) aplican técnicas de ML para optimizar un portafolio de inversiones con acciones del índice COLCAP, seleccionando finalmente las acciones de Ecopetrol, Bancolombia y Celsia, utilizando como periodo de tiempo 2018-2022. Al igual que los trabajos anteriores, se busca predecir la dirección del retorno diario. Los resultados determinaron que el modelo Random Forest ofrece el mejor desempeño, proporcionando un portafolio con una rentabilidad de 18.96% y una desviación estándar de 20.60%, generando así un ratio de Sharpe de 0.40, valor bastante similar al obtenido en el portafolio construido bajo la metodología de Markowitz (0.41).

Los trabajos presentados anteriormente revelan algunas consideraciones importantes que requieren un análisis crítico. Primero, resultados excesivamente optimistas como los presentados por Gumelar et al. (2020), con precisiones cercanas al 100%, llevan a cuestionar su validez y sugieren posibles casos de backtest overfitting, tal como lo plantea López de Prado (2018). Resultados tan positivos seguramente no sean replicables en condiciones reales de mercado.

Segundo, existe una tendencia generalizada hacia horizontes de predicción de corto plazo (de frecuencia diaria o semanal), lo cual lleva a que los portafolios construidos con estas metodologías estén enfocados en aprovechar oportunidades de inversión a corto plazo.

### ***6.1.3 Optimización de asignación de capital***

La optimización de la asignación de capital representa un avance lógico desde la predicción hacia la construcción directa de portafolios.

López de Prado (2016) plantea la metodología “Hierarchical Risk Parity” (HRP), la cual tiene como base la teoría de gráficos y machine learning. Este busca solucionar tres problemas principales de la metodología de Markowitz: inestabilidad, concentración en las asignaciones y un pobre desempeño por fuera de la muestra. Específicamente este procedimiento permite aprovechar en mayor medida la información disponible en la matriz de covarianzas al construir una estructura jerárquica que agrupa los activos en diferentes clústeres según su similitud, lo cual permite al portafolio mantener la diversificación entre cada uno de los clústeres, pero realizar los ajustes necesarios en las ponderaciones dentro de cada cluster. Al realizar una comprobación mediante simulaciones de Monte Carlo, fue posible determinar que el portafolio construido bajo HRP logra obtener una varianza 72% menor que el portafolio construido bajo MVO, lo cual lleva a una mejora en el ratio de Sharpe de 31.3%.

Jang & Seong (2023) buscan generar un enfoque innovador al conectar la teoría moderna de portafolios (MPT) con técnicas de ML, específicamente Deep Reinforcement (DRL). Los autores implementan una estrategia que busca distribuir el capital entre los diferentes activos teniendo en cuenta su información histórica (principalmente su correlación), así como las señales generadas por los algoritmos al analizar indicadores de análisis técnico. Lo anterior permite que, en comparación con métodos como MVO, donde las asignaciones suelen ser poco dinámicas (rebalanceos anuales), el método propuesto logra ajustar dinámicamente las asignaciones de capital dadas las condiciones del mercado. En términos prácticos, tras evaluar el modelo con 29 acciones del índice Dow Jones Industrial Average (DJIA) en el periodo de tiempo 2008-2019, fue

posible determinar que este logra superar notablemente a otros algoritmos, obteniendo un ratio de Sharpe de 2.0067, mientras que el segundo mejor modelo logró un ratio de 1.6838.

Zhang et al. (2021) proponen un framework universal “End-to-End” para la optimización de portafolios que busca superar las limitaciones del enfoque tradicional de optimización media-varianza propuesto por Markowitz mediante el uso de algoritmos de deep learning. Su investigación aborda principalmente tres problemas: la dificultad de predecir los retornos futuros, la inestabilidad al estimar la matriz de covarianzas y la incorporación de restricciones prácticas.

Su propuesta elimina completamente la necesidad de generar predicciones de retornos al modelar directamente las ponderaciones en el portafolio a través de redes neuronales profundas. La arquitectura propuesta se compone de dos bloques: un bloque de puntuación (score block) que transforma los datos en puntuaciones para cada acción, y un bloque de portafolio (portfolio block) que convierte dichas puntuaciones en ponderaciones que satisfacen las restricciones impuestas.

El trabajo de Zhang et al. (2021) resulta innovador no únicamente por eliminar la necesidad de generar predicciones de retornos, sino también al manejar las restricciones impuestas a través de capas neuronales diferenciables. Estas restricciones corresponden específicamente al número de acciones en el portafolio, la ponderación máxima de cada acción, las ventas en corto y apalancamiento.

La validación de resultados es realizada en dos conjuntos de datos: uno sintético y otro real con 735 acciones del índice Russell 3000 cuyos datos corresponden a los años 1984-2021, siendo el periodo de prueba los años 2001-2021. En el conjunto de datos real, el modelo E2E-LSTM logra un ratio de Sharpe de 2.604, comparado con 0.858 del mejor modelo tradicional (CS-LSTM), representando una mejora significativa. Adicionalmente, al observar el desempeño durante

periodos de crisis financiera como 2008 y la pandemia de 2020, el portafolio experimenta un comportamiento robusto.

El trabajo de Zhang et al. (2021) establece un precedente metodológico significativo al demostrar la posibilidad de incluir directamente restricciones complejas en las arquitecturas de los modelos de aprendizaje profundo, logrando así posicionarse como una herramienta versátil para la gestión de portafolios. Sus resultados respaldan la superioridad de algoritmos de aprendizaje profundo respecto a las técnicas tradicionales de optimización, especialmente en su capacidad para manejar múltiples funciones objetivo y restricciones sin afectar su desempeño.

#### ***6.1.4 Revisiones sistemáticas de la literatura***

La creciente aplicación de técnicas de IA en la gestión de portafolios ha motivado el desarrollo de revisiones sistemáticas que buscan consolidar el conocimiento existente e identificar tendencias en la investigación académica.

Gao et al. (2024) desarrollan una importante revisión sistemática que proporciona una perspectiva comprehensiva sobre las múltiples aplicaciones del machine learning en los negocios y las finanzas. Tras identificar 68 artículos publicados en bases de datos académicas entre 2007-2021, los autores evidencian una clara preferencia por los modelos de RNN (20 estudios), SVM (18 estudios) y Random Forest (17 estudios) al ser los más utilizados, específicamente en las tareas relacionadas mayormente con el mercado de valores.

Los autores resaltan la gran importancia que han tenido a lo largo de los años los datos cuantitativos, sin embargo, los recientes avances en los modelos de IA permiten considerar fuentes alternativas de datos, específicamente datos cualitativos como pueden serlo las redes sociales, informes y demás fuentes textuales que podrían aportar información adicional al análisis. Adicionalmente, los autores presentan la discusión referente a la difícil interpretabilidad de

resultados característica de los modelos más complejos, lo cual puede suponer serios problemas al momento de justificar o tomar decisiones que deban obedecer a estrictas regulaciones o procedimientos que garanticen su idoneidad.

Los artículos analizados por los autores denotan una clara preferencia por estudiar dinámicas de los mercados de países desarrollados, como Estados Unidos, Reino Unido, Japón y Corea del Sur, mientras que países emergentes cuentan con una limitada evidencia empírica. Todo lo anterior contribuye a evidenciar la relevancia que tienen investigaciones como esta al confirmar que la aplicación de técnicas de ML puede llevar a mejores resultados financieros en áreas como la gestión de portafolios, así como refleja la necesidad de llevar a cabo mayor investigación en mercados emergentes.

De forma similar a la investigación llevada a cabo por Gao et al. (2024), Sutiene et al. (2024) realizan una revisión de literatura donde analizan una variedad de trabajos que buscan presentar las contribuciones de la IA a las diferentes etapas de la gestión de portafolios, siendo estas la planificación, ejecución y retroalimentación.

Específicamente, Sutiene et al. (2024) demuestran que la evolución tecnológica hacia modelos más complejos y con mayores capacidades, como las redes neuronales LSTM, lleva a generar resultados superiores respecto a métodos tradicionales en tareas de predicción de series de tiempo financieras, gracias a sus capacidades de procesar mayores volúmenes de datos y capturar relaciones complejas entre datos secuenciales.

Los autores presentan una importante discusión respecto a la inteligencia artificial explicable (XAI) al considerarla como un requisito indispensable para su adopción en masa, sobre todo en entornos institucionales. La tendencia a crear modelos cada vez más complejos lleva a

enfrentar el problema de la “caja negra”, comprometiendo así la transparencia necesaria para llevar a cabo decisiones de inversión, especialmente en contextos donde existen regulaciones estrictas.

Se identifica como un factor diferenciador la capacidad de los algoritmos para implementar estrategias de rebalanceo dinámico mediante técnicas como el aprendizaje por refuerzo, que respondan a las condiciones del mercado en tiempo real, superando las limitaciones que enfrentan las estrategias tradicionales al realizar rebalanceos en fechas específicas (trimestral, semi anual, anual) o basados en umbrales fijos.

Sutiene et al. (2024) reconocen que la implementación exitosa de sistemas de IA en la gestión de portafolios requiere un abordaje multidimensional, que considere a la vez su desempeño financiero, la interpretabilidad de los algoritmos y el cumplimiento regulatorio. Por tanto, la gestión de portafolios basada en IA requiere ser vista como una disciplina que necesita metodologías robustas para validar su importancia en contextos como el mercado colombiano.

## **6.2 Bolsa de Valores de Colombia**

La presente sección busca realizar una breve descripción de las principales acciones que conforman la Bolsa de Valores de Colombia, pues teniendo en cuenta el objetivo de la presente investigación, resulta necesario tener una contextualización y comprensión de las empresas cotizadas. En la actualidad el mercado bursátil colombiano cuenta con aproximadamente 30 acciones listadas, las cuales pertenecen a seis sectores claves: servicios financieros, servicios públicos, energía, industria y bienes y servicios de consumo. Sin embargo, dados los criterios de selección no todas estas pudieron ser incluidas en la presente investigación dado que la información bursátil para algunas de ellas resulta ser bastante limitada dado su poco tiempo presentes en la BVC.

A continuación, se ofrecerá una descripción general de las empresas que componen cada sector.

### **6.2.1 Servicios financieros**

**Bancolombia SA:** Bancolombia es un grupo financiero fundado el 29 de enero de 1875 en la ciudad de Medellín, contando hasta el momento con una experiencia de 150 años en la industria financiera. A lo largo de su historia ha logrado expandir su alcance más allá de Colombia, llegando a países como Panamá, Guatemala y El Salvador, empleando en su totalidad a cerca de 30000 empleados. Adicionalmente conviene destacar el hecho de que el 25 de julio de 1995 se convirtió en la primera entidad financiera del país en ser listada en la bolsa de valores de Nueva York (NYSE) (Grupo Bancolombia, 2025).

**Grupo de Inversiones Suramericana:** El grupo Sura se define a sí mismo como un gestor de inversiones enfocado en aprovechar las oportunidades de América Latina en sectores como los seguros, ahorros, administración de activos y banca. Desde su creación el 12 de diciembre de 1944 en la ciudad de Medellín hasta la actualidad, el grupo Sura ha experimentado un sostenido crecimiento que le ha permitido estar presente en 11 países de la región. Las actividades de este grupo se encuentran concentradas principalmente en tres organizaciones: 1) Suramericana, la cual es su organización especializada en seguros y cuyo alcance la convierte en la más grande de América Latina; 2) SURA Asset Management, corresponde a su filial especializada en ahorro, inversión y gestión de activos; 3) Bancolombia, empresa de la cual es propietaria del 46.11% de las acciones ordinarias (Grupo Sura, 2024).

**Corporación Financiera Colombiana:** Desde su creación en 1961 Corficolombiana ha desempeñado un rol de liderazgo en la promoción de inversiones y proyectos enfocadas en diferentes sectores clave de la economía del país, en especial la agroindustria, hostelería,

infraestructura y energía y gas. Con fecha de corte a junio de 2024, Corficolombiana contaba con más de 26000 trabajadores y activos por valor de 7000 millones de dólares (Corficolombiana, 2025).

Grupo Bolívar: Constituido el 5 de diciembre de 1939 en la ciudad de Bogotá, el Grupo Bolívar ha evolucionado hasta convertirse en uno de los mayores grupos empresariales del país, al contar con empresas del sector financiero, sector asegurador e incluso del sector de la construcción y desarrollo inmobiliario (Grupo Bolívar, 2012).

Banco de Bogotá: Fundado en el año 1870, es uno de los bancos con mayor tiempo de funcionamiento y experiencia en el país, razón por la cual actualmente goza de una posición de liderazgo en el sector financiero. Así mismo, es una empresa filial del importante Grupo Aval (Banco de Bogotá, 2013).

Bolsa de Valores de Colombia: Creada en julio de 2001 luego de la unión de las bolsas de Bogotá, Medellín y Occidente, la BVC se ha convertido en un aliado estratégico del sector financiero y bursátil en Colombia al estar a cargo de la operación del mercado de capitales (BVC, 2025).

### **6.2.2 Servicios Públicos**

Interconexión Eléctrica SA: Desde su creación en la ciudad de Bogotá en 1967, ISA ha logrado ocupar una posición de liderazgo en el sector de la energía eléctrica no solo en el país, sino también en la región al contar con presencia en Panamá, Perú, Brasil, Bolivia, Chile y Centroamérica, empleando un total de más de 5000 personas. De igual forma, sus operaciones se extienden a las vías (Mayormente en Chile) y a las TICS (ISA, 2025).

Grupo Energía Bogotá SA: Gracias al ser constituida en 1896 fue una empresa pionera en la generación y distribución de energía eléctrica en el país. Desde aquel momento logró mantener

una senda de crecimiento positivo hasta la actualidad, momento en el cual cuenta con operaciones a lo largo de América Latina en las diferentes etapas de generación, transmisión y distribución. Así mismo, en décadas recientes ha incursionado en el negocio del transporte y distribución de gas natural. Actualmente cuenta con más de 2300 trabajadores (GEB, 2025).

**Celsia ESP SA:** Celsia es una empresa encargada de la generación, transmisión y comercialización de energía proveniente de fuentes renovables, llevando a cabo operaciones en Colombia, Panamá y Honduras, disponiendo de un total de 40 centrales de generación. En Colombia cuenta con 1.3 millones de clientes pertenecientes a los departamentos del Valle del Cauca y Tolima. Cabe resaltar que Celsia es la empresa de energía del Grupo Argos (Celsia, 2024).

**Promigas SA:** Creada en el año 1974 en la ciudad de Barranquilla, Promigas ha tenido un rol importante en la conformación de la industria del transporte y distribución de gas natural en Colombia, y más recientemente en la regasificación de GNL y distribución de energía eléctrica. En Colombia, Promigas cuenta con una cuota del 38% del mercado de distribución de gas natural, mientras que en Perú esta cifra asciende al 96% del mercado (Promigas, 2024).

### ***6.2.3 Energía***

**Ecopetrol SA:** Desde su creación el 25 de agosto de 1951 Ecopetrol ha desempeñado un rol muy importante para la economía de Colombia, tanto así que durante muchos años ha sido la empresa más grande del país. Tiene operaciones en la totalidad de los eslabones del sector petrolero, disponiendo de pozos de extracción, refinerías, transporte y logística de hidrocarburos, así como oleoductos y poliductos. Sus operaciones no se limitan únicamente a Colombia, pues estas llegan a estar presentes en Brasil, México y Estados Unidos (Ecopetrol, 2021).

**Organización Terpel SA:** Desde su creación en 1968 en la ciudad de Bucaramanga con una cantidad inicial de 20 estaciones de servicio, Terpel ha logrado convertirse en uno de los

principales actores en el mercado de los combustibles en el país, llegando a contar con operaciones en República Dominicana, Panamá, Ecuador y Perú. Actualmente Terpel dispone de más de 2400 estaciones de servicio en Latinoamérica, a la vez que emplea de forma directa a 3400 trabajadores y 8000 de forma indirecta (Terpel, 2025).

Canacol Energy LTD: Siendo de origen canadiense, desde el año 2012 Canacol Energy ha redirigido la totalidad de su negocio desde la extracción petrolera hasta la exploración y extracción de gas natural on-shore, esto como parte de una estrategia para la reducción de gases de efecto invernadero, logrando en la actualidad ser la empresa más grande de su tipo en el país, al abastecer aproximadamente el 17% de la demanda nacional (Canacol Energy Ltd, 2025).

#### **6.2.4 Industria**

Cementos Argos SA: Establecida el 27 de febrero de 1934 en la ciudad de Medellín, Cementos Argos ha contado con una destacada trayectoria empresarial sirviendo de fundamento para el desarrollo del país. Actualmente está presente en 16 países, siendo posible destacar su posición de liderazgo al ser la mayor cementera y concretera de Colombia, así como su importante posicionamiento en Centroamérica. Así mismo está presente en el mercado americano a través de Summit Materials, empresa de la cual posee el 31%. Con corte a 2023, Argos cuenta con cerca de 4000 empleados (Argos, 2024).

Inversiones Argos SA: El Grupo Argos constituye en la actualidad un holding empresarial enfocado en la infraestructura e industria cementera, específicamente sus filiales son Cementos Argos, Celsia, Odinsa y cierta línea de negocio enfocada al desarrollo urbano. En su totalidad, el Grupo Argos cuenta con más de 11000 trabajadores (Grupo Argos, 2024).

Constructora Conconcreto SA: Fundada en el año 1961 por parte de un grupo de arquitectos e ingenieros, Conconcreto es una importante empresa dedicada al diseño, construcción y gerencia

de proyectos de infraestructura para diversos sectores como el aeroportuario, logístico, comercial, hotelero, industrial, transporte y de la salud. Su desempeño le ha permitido internacionalizarse y contar con presencia en Panamá y en la Florida en Estados Unidos (Concreto, 2024).

### **6.2.5 Servicios y bienes de consumo**

Almacenes Éxito SA: Remontando su historia al año 1905 con la apertura de Carulla en Bogotá, al año 1992 con el inicio de LEY en Barranquilla y en marzo de 1949 en Medellín con los inicios de Éxito, el Grupo Éxito surgió como una tienda enfocada en los supermercados y al negocio textil, la cual posteriormente iniciaría un proceso de expansión hasta lograr convertirse en la actualidad en una de las mayores cadenas de supermercados del país, llegando incluso a contar con una importante presencia en Uruguay y Argentina. En su totalidad el Grupo Éxito cuenta con más de 43000 empleados (Grupo Éxito, 2025).

Grupo Nutresa SA: Los inicios de esta empresa se remontan al 12 de abril de 1920 en el municipio de Sonsón, Antioquia, donde fue constituida inicialmente bajo el nombre de Compañía Nacional de Chocolates Cruz Roja. Desde entonces ha iniciado un constante proceso de expansión en búsqueda de adquirir nuevas empresas y líneas de negocio enfocadas en la industria alimentaria. Su gran capacidad le ha permitido estar presente en 6 países de la región, mientras que sus productos son comercializados en un total de 18 países (Grupo Nutresa, 2022).

Fabricato SA: Su historia se remonta al año 1919 en el municipio de Bello, Antioquia, año desde el cual ha estado enfocada en la fabricación de textiles, logrando en la actualidad posicionarse como una de las más importantes del país. Con fecha de corte a 2024 disponía de un total de 1100 trabajadores (Fabricato, 2024).

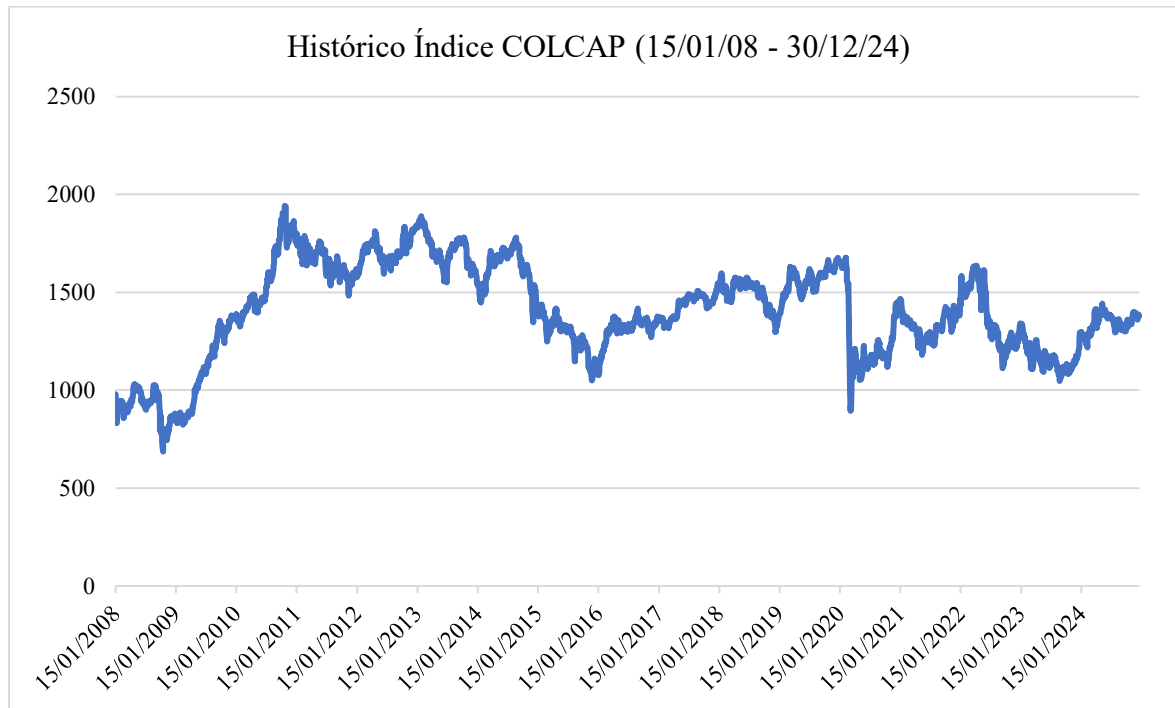
### **6.2.6 Materiales básicos**

Mineros SA: Es una empresa con más de 50 años de trayectoria en la exploración y extracción de oro cuyas oficinas centrales están ubicadas en Medellín. Sus operaciones están concentradas principalmente en Colombia y Nicaragua, aunque cuenta con participación en proyectos de exploración en Chile. Al cierre del año 2023, Mineros contaba con 2373 empleados.

### **6.3 Índice de referencia COLCAP**

Las diferentes bolsas de valores alrededor del mundo tienen a disposición de los inversores una amplia variedad de acciones en las cuales invertir. Sin embargo, con el objetivo de observar la dinámica de los mercados financieros son elaborados ciertos índices de referencia que buscan reflejar el comportamiento agregado y ponderado de los activos cotizados, siendo quizás los más reconocidos a nivel mundial el S&P 500 (Este refleja el comportamiento de las 500 empresas más grandes de los Estados Unidos) y el DJI (30 empresas con mayor capitalización de los Estados Unidos).

En el caso de Colombia, originalmente el índice de referencia recibió el nombre de Índice General de la Bolsa de Valores de Colombia (IGBC), siendo calculado por primera vez el 15 de enero de 2008, contando con un valor inicial de 1000 puntos. Hacia el año 2013, específicamente el 1 de noviembre, el IGBC presentó un cambio en su nombre pasándose a llamar COLCAP, nombre que mantiene en la actualidad. Este índice refleja el comportamiento de las 20 acciones con mayor liquidez de la BVC, ajustando la ponderación de acuerdo con su capitalización bursátil.

**Figura 5. Comportamiento histórico índice COLCAP**

Nota: Elaboración propia con base en datos del Banco de la República

La información presentada en la anterior gráfica muestra el comportamiento histórico del índice COLCAP desde el primer momento en que fue calculado (15/01/08) hasta finales del año 2024. Durante los primeros años de su existencia fue influido negativamente por los impactos generados por la crisis económica de 2008, explicando así su caída hasta alcanzar mínimos históricos de 686.65 puntos el 27 de octubre de 2008. Desde ese punto inició un periodo de tiempo caracterizado por su rápido y constante ascenso, llegando a alcanzar su máximo histórico de 1942.37 puntos el 4 de noviembre de 2010. Sin embargo, desde entonces el comportamiento de este índice ha estado marcado por una tendencia lateral, generando así muy pocas oportunidades de inversión a largo plazo para los inversionistas.

Dando un salto hasta el año 2020 es importante mencionar el impacto que tuvo la crisis económica causada por el COVID-19, pues cómo es posible apreciar el índice COLCAP

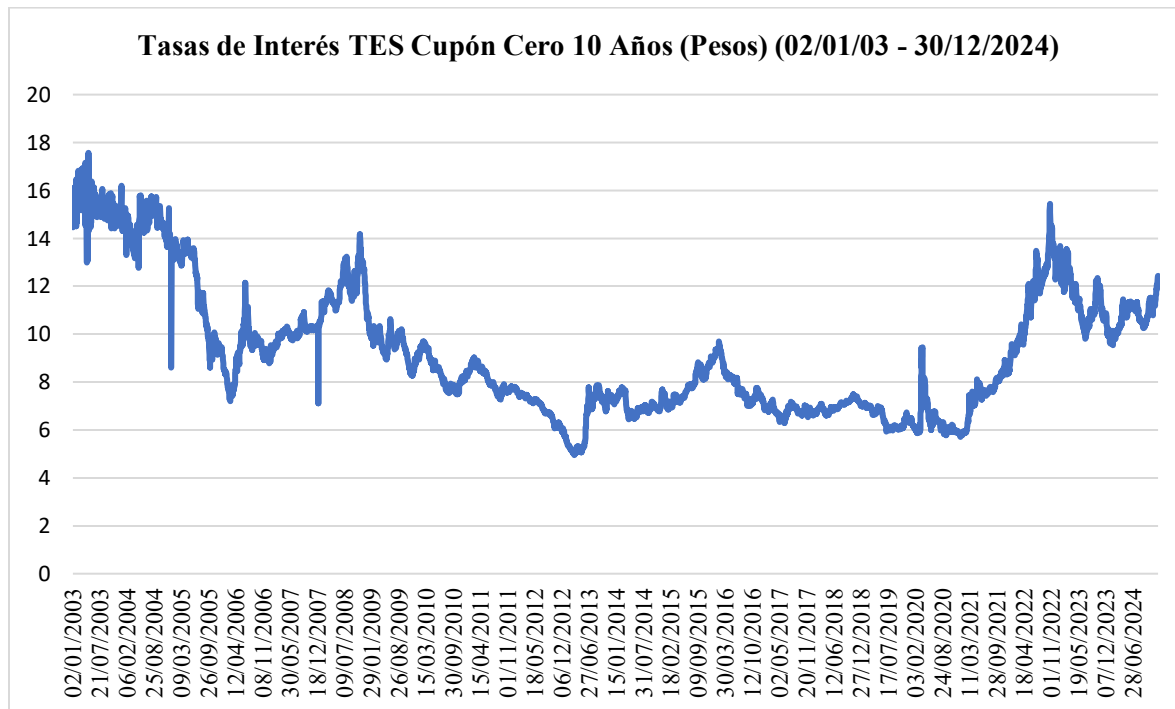
experimentó una dramática caída del 46.67%, esto considerando su valor máximo a principios de 2020 (1676.49 puntos el 3/01/2020) respecto a su punto más bajo después de ser declarada la pandemia (894.03 puntos el 18/03/2020).

Ahora, teniendo en cuenta el comportamiento histórico de este índice, resulta conveniente realizar unos simples cálculos para determinar la rentabilidad obtenida por un inversor el cual haya decidido invertir desde el momento de su creación hasta finales del año 2024. Como es de esperar, los resultados son bastante pobres. Tal inversor tan solo hubiera obtenido una rentabilidad del 40.74%, o en términos anuales una rentabilidad de tan solo 2.16%. Esto debe ser puesto en perspectiva. Tradicionalmente se argumenta desde el punto de vista de las finanzas que aquellas clases de activos con mayor riesgo normalmente suelen ser compensados con una mayor rentabilidad, siendo el ejemplo más utilizado la comparación entre los bonos y las acciones. Mientras que los primeros suelen tener el respaldo de importantes instituciones como el Estado o empresas, así como la característica de ofrecer pagos periódicos por montos predefinidos, las acciones constituyen una reclamación residual sobre las utilidades generadas por la empresa una vez sean saldadas sus obligaciones financieras, sin dejar de lado el hecho que sus cotizaciones pueden experimentar fluctuaciones significativas. Técnicamente recibe el nombre de prima de riesgo de las acciones (*Equity Risk Premium*), siendo así el exceso de rentabilidad que se espera obtener al invertir en acciones en vez de activos más seguros.

Al observar el comportamiento de las tasas de interés de los Títulos de Tesorería (TES) a 10 años (figura 6), los cuales en el caso del mercado colombiano pueden ser considerados como el activo libre de riesgo, surgen algunas consideraciones importantes de cara a la inversión en acciones. Partiendo de la rentabilidad anual lograda por el COLCAP a lo largo de su historia (2.16%), los TES a 10 años han brindado consistentemente rentabilidades superiores a esta, de

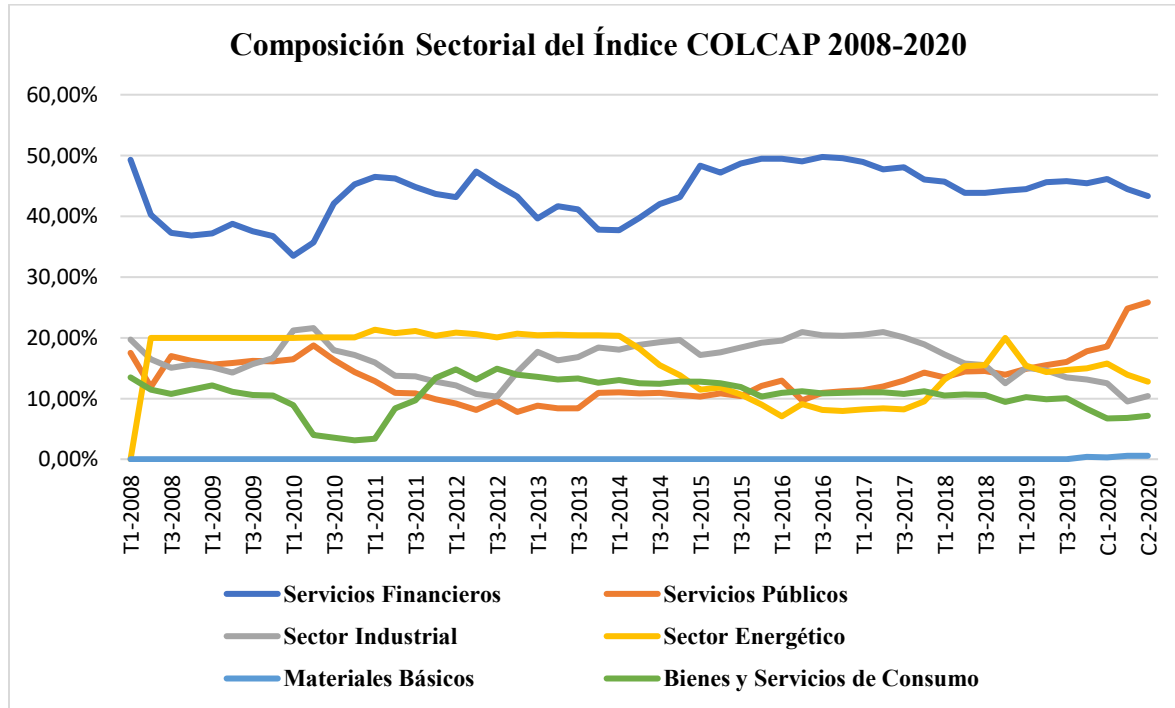
hecho, al calcular su tasa de interés promedio durante el mismo periodo de tiempo que el COLCAP este ha sido de 8.35%. Esto plantea una contracción importante respecto a los planteamientos de la teoría financiera ya que los inversores, contrario a obtener una prima de riesgo de las acciones positiva, han estado obteniendo una prima negativa. Es decir, más allá de ser compensados por invertir en acciones, han sido castigados por invertir en acciones.

**Figura 6. Tasas de interés históricas TES 10 años**



Nota: Elaboración propia con base en datos del Banco de la República

Dada la dinámica que pueden tener las acciones pertenecientes a los diferentes sectores que conforman el índice es conveniente observar la composición histórica de cada uno de estos.

**Figura 7. Composición sectorial histórica índice COLCAP**

Nota: Elaboración propia con base en datos de Bolsa de Valores de Colombia

De la figura 7 es posible apreciar ciertas características del COLCAP. Primero, es un índice compuesto principalmente por acciones pertenecientes al sector financiero, cuya participación ha rondado entre el 33-49%, siendo su menor participación alcanzada en años posteriores a la crisis financiera del 2008. Por otro lado, el sector energético que históricamente ha resultado ser de gran importancia para el país refleja una dinámica bastante relacionada con los precios internacionales del petróleo, esto al reducir su participación en el índice tras la caída del precio del petróleo que tuvo lugar en 2014.

Mientras tanto, sectores como el industrial, de servicios públicos y de bienes y servicios de consumo han presentado participaciones que rondan generalmente entre el 7 y 20%. En este aspecto es necesario mencionar el repunte que tuvieron los servicios públicos, este tuvo lugar dada la naturaleza misma de estas empresas, las cuales a pesar de las medidas tomadas como respuesta

a la pandemia del COVID-19 pudieron mantener su funcionamiento y generación de ingresos con una menor afectación relativa a las otras empresas. Finalmente, el sector de materiales básicos está conformado por una única acción, Mineros, la cual logra entrar al índice COLCAP tal solo hasta el cuarto trimestre del año 2019, de allí la razón de que durante la mayor parte de tiempo su participación haya sido nula.

#### **6.4 Portafolios metodología de Markowitz**

En la presente sección son presentados los resultados obtenidos tras haber empleado la metodología media-varianza propuesta por Harry Markowitz (1952) para la optimización de portafolios, en este caso para las empresas listadas en la Bolsa de Valores de Colombia. El periodo de tiempo utilizado corresponde a 6 años, de 2019 a 2024, siendo el periodo de tiempo 2019-2023 utilizado para calcular la matriz de covarianzas y retornos promedios, mientras que los datos del año 2024 fueron utilizados para evaluar el desempeño del portafolio. Como tasa libre de riesgo fue utilizada la tasa promedio para los TES a 10 años en pesos colombianos (COP) durante el periodo 2019-2023, siendo equivalente a 8.68%.

Así mismo otro de los parámetros para tener en cuenta en la respectiva optimización corresponde a la ponderación máxima que puede tener alguna acción en el portafolio, cuyo parámetro se definió en 20%. Respecto a las operaciones de compra y venta, se definió que el portafolio únicamente consistirá en posiciones en largo, es decir, ningún activo podrá tener una ponderación negativa (ventas en corto).

Mediante la siguiente fórmula, fue calculada la rentabilidad logarítmica anual para cada una de las acciones listadas en la BVC:

$$R_{anual} = \left[ \exp \left( \left( \frac{1}{n} \right) * \sum \ln \left( \frac{P_t}{P_{t-1}} \right) \right) - 1 \right] \times d$$

Donde:

- $R_{anual}$  corresponde a la rentabilidad anual
- $n$  es el número total de observaciones
- $P_t$  corresponde al precio de la acción en el día  $t$
- $P_{t-1}$  corresponde al precio de la acción en el día  $t-1$
- $d$  es el número de días bursátiles en el año (241 días)

**Tabla 7. Rentabilidades anuales periodo 2019 - 2023**

<b>Ticker</b>	<b>Rentabilidad Anual (2019-2023)</b>
BCOLOMBIA	6,55%
BOGOTA	-5,67%
BVC	6,11%
CELSIA	-0,07%
CEMARGOS	2,12%
CNEC	-16,86%
CONCONCRET	-5,62%
CORFICOLCF	6,71%
ECOPETROL	5,96%
ÉXITO	2,91%
FABRICATO	-5,64%
GEB	9,52%
GRUBOLIVAR	5,37%
GRUPOARGOS	-2,47%
GRUPOSURA	0,43%
ISA	6,23%
MINEROS	0,64%
NUTRESA	17,22%
PROMIGAS	1,45%
TERPEL	2,38%

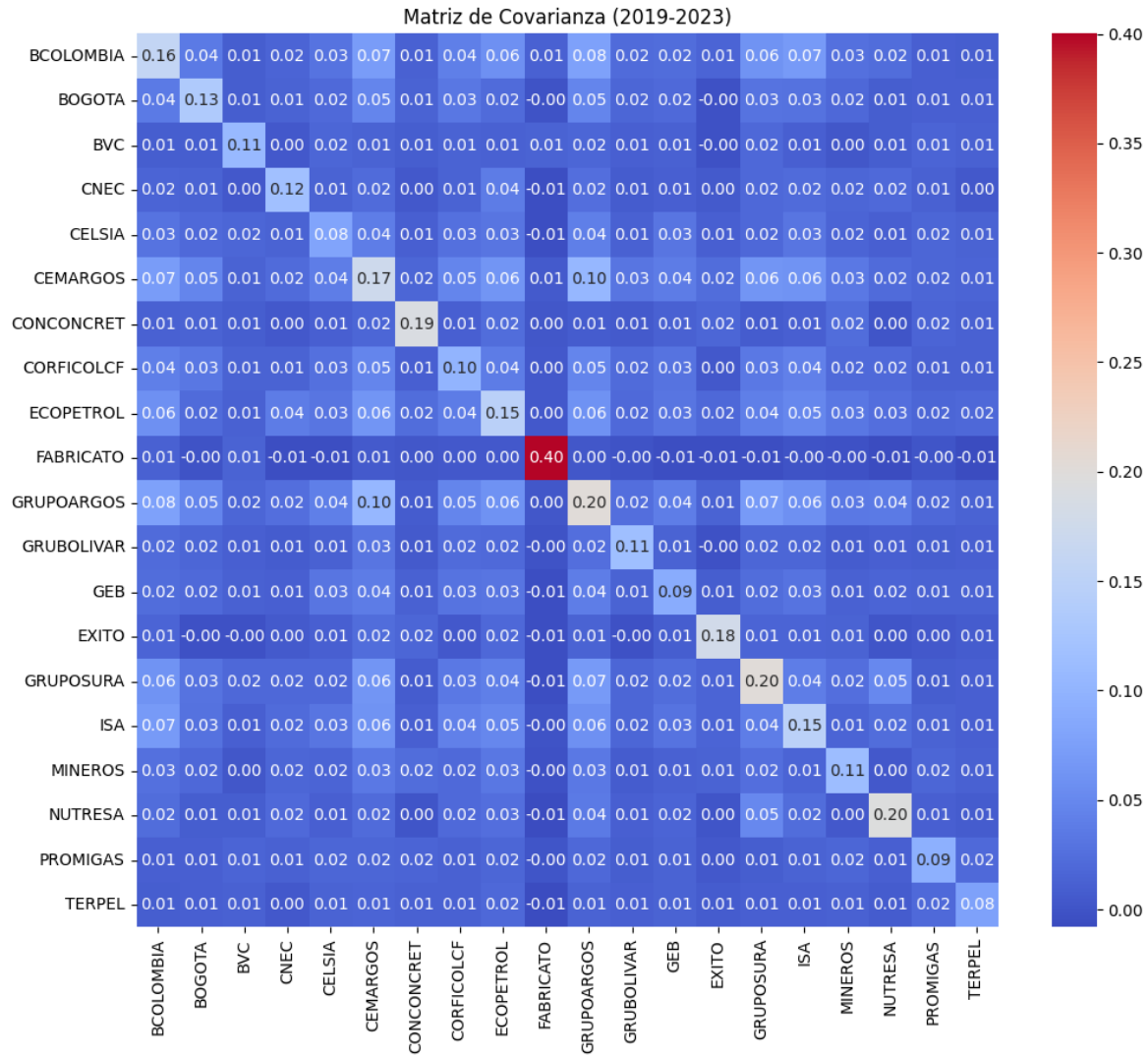
Nota: Elaboración propia con base en datos de Yahoo Finance

De la anterior tabla es posible destacar el comportamiento positivo que han tenido las acciones de Nutresa (17.22%) y GEB (9.52%). Por el contrario, las acciones de CNEC (-16.86%), BOGOTA (-5.67%), FABRICATO (-5.64%), CONCRET (-5.62%) y GRUPOARGOS (-2.47%) han presentado comportamientos negativos que reflejan las dinámicas que han experimentado sectores como el energético y constructor durante los últimos años.

Los anteriores retornos anuales estimados serán utilizados como la rentabilidad esperada para cada una de las acciones, información que es necesaria disponer para llevar a cabo la optimización del portafolio, específicamente para aquel que busque maximizar el ratio de Sharpe.

Continuando con las estimaciones necesarias, fue calculada la matriz de covarianzas entre las diferentes acciones partiendo de los retornos logarítmicos diarios, siendo esta información presentada en el siguiente mapa de calor.

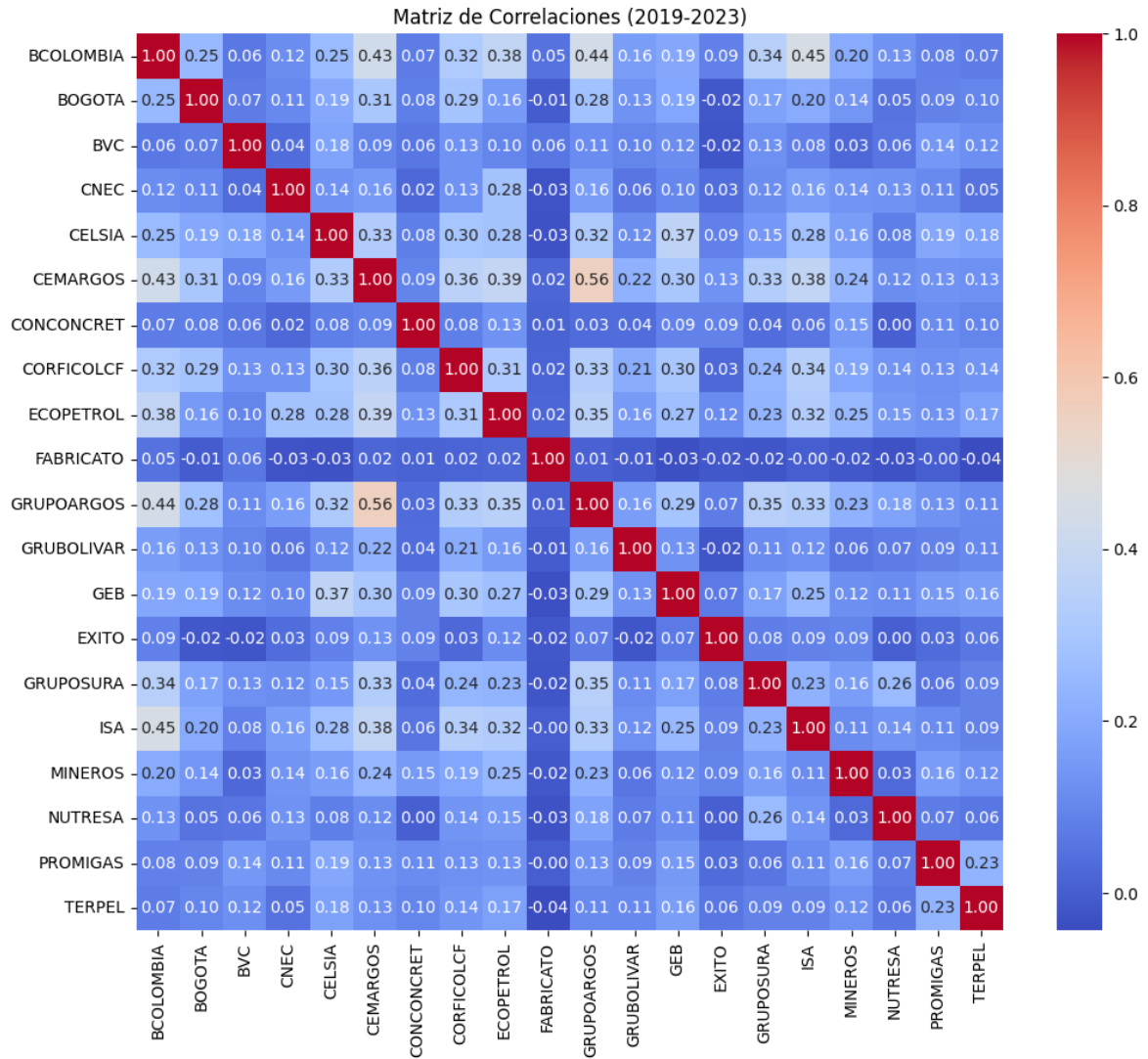
**Tabla 8. Matriz de Covarianzas Anualizada (2019-2023)**



Nota: Elaboración propia con base en datos de Yahoo Finance

Adicionalmente, para facilitar la interpretación de la relación presente entre las diferentes acciones fue calculada la matriz de correlaciones, cuya interpretación resulta más intuitiva.

**Tabla 9. Matriz de Correlaciones (2019-2023)**



Nota: Elaboración propia con base en datos de Yahoo Finance

De la anterior matriz de correlaciones es posible destacar algunos hechos. La mayor parte de las acciones presentan correlaciones positivas entre si (ya sean débiles o fuertes), lo cual reduce de cierta manera la capacidad del portafolio para llevar a cabo una diversificación estratégica que permita absorber o contrarrestar de mejor forma las variaciones negativas de las demás acciones. Específicamente, las acciones de FABRICATO y ÉXITO son aquellas que tienden a experimentar

una correlación negativa con las demás acciones, motivo por el cual resulta más probable que estas sean incluidas en el portafolio al momento de optimizarlo.

Por otra parte, aunque la mayoría de las acciones cuenten con correlaciones positivas estas no suelen ser mayores a 0.5, dejando un margen considerable para aprovechar los beneficios de la diversificación, ya que tal como mencionó Markowitz (1952), siempre que la correlación sea diferente a 1, será posible reducir la volatilidad del portafolio a través de la diversificación. Finalmente, resulta conveniente aclarar que la correlación de 0.56 entre GRUPOARGOS y CEMARGOS tiene su origen en el hecho de que GRUPOARGOS es el accionista mayoritario de CEMARGOS.

#### **6.4.1 Optimización de Portafolios y Frontera Eficiente**

Con base en la información presentada hasta el momento, se implementó la optimización del portafolio de Markowitz y el de mínima varianza, así como la posterior representación gráfica de estos portafolios en la frontera eficiente.

Los resultados obtenidos mediante la aplicación de técnicas de optimización, en este caso de la optimización media-varianza (MVO), son presentados en la tabla 10, donde se evidencian diferencias significativas entre ambos portafolios.

**Tabla 10. Retorno y Volatilidad Esperados**

<b>Portafolio</b>	<b>Retorno</b>	<b>Volatilidad</b>	<b>Ratio de Sharpe</b>
Markowitz	9,23%	20,10%	0.0565
Mínima Varianza	1,67%	13.57%	-0.5169

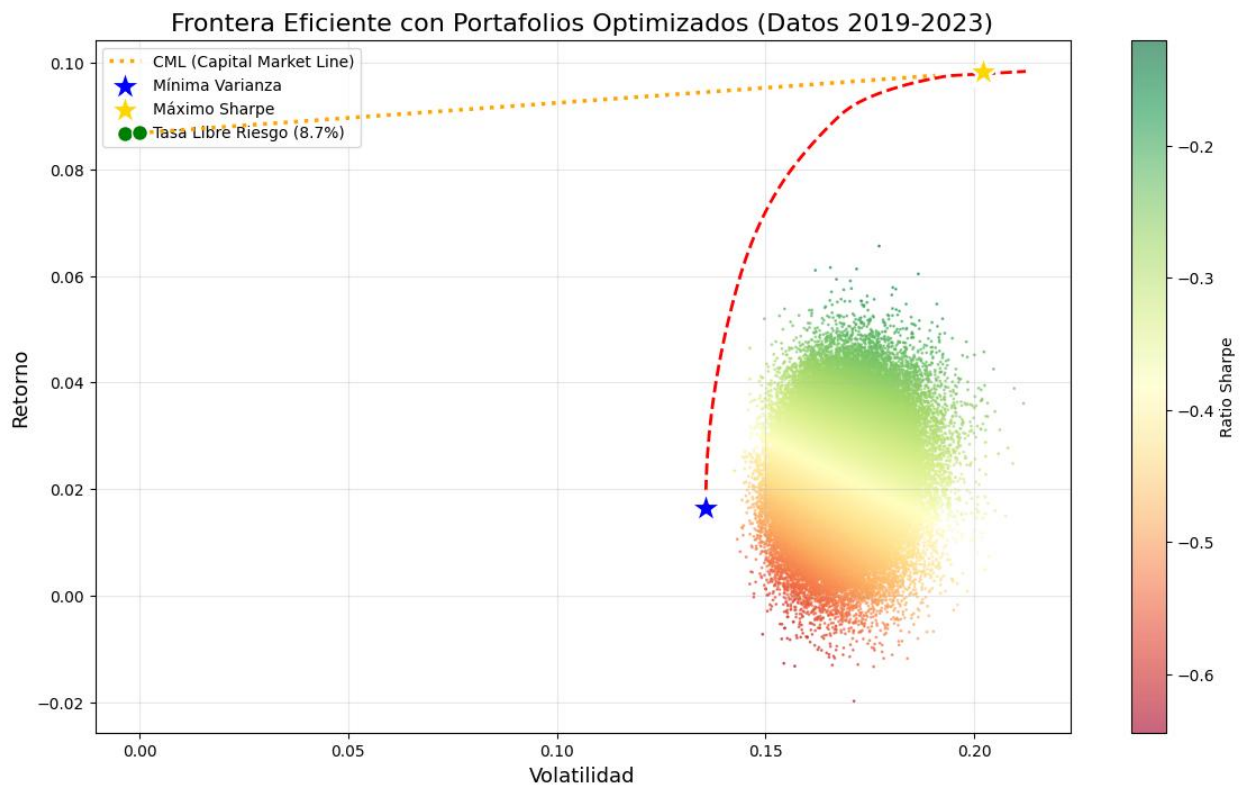
Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

El portafolio de Markowitz alcanzó un retorno esperado anualizado del 9.23% con una volatilidad del 20.10%, generando un ratio de Sharpe positivo de 0.0565. este resultado indica que el portafolio logró una compensación adecuada por el riesgo asumido, superando la tasa libre de riesgo durante el periodo analizado (2019-2023).

Por el contrario, el portafolio de mínima varianza, si bien logró reducir considerablemente la volatilidad hasta el 13.57%, obtuvo un rendimiento esperado de apenas 1.67%, resultando en un ratio de Sharpe negativo de -0.5169. Este desempeño sugiere que, para el periodo 2019-2023, la estrategia de minimizar al máximo el riesgo no brindó una adecuada compensación para el inversor, situándose por debajo del rendimiento de la tasa libre de riesgo (8.7%).

La figura 8 muestra la implementación de la frontera eficiente, resaltando de igual forma la línea del mercado de capitales (CML por sus siglas en inglés), el portafolio de mínima varianza y el portafolio de Markowitz (máximo Sharpe).

**Figura 8. Frontera Eficiente**



Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

La frontera eficiente (línea roja) representa todos aquellos puntos de portafolios eficientes, en el sentido de que maximizan el retorno dado cierto nivel de riesgo, o minimizan el riesgo dado

cierto nivel de retorno. Todo portafolio ubicado en la frontera eficiente es óptimo, mientras que los portafolios ubicados dentro de la frontera eficiente son factibles, pero no son eficientes.

La línea amarilla representa la línea del mercado de capitales (CML), que se extiende desde la tasa libre de riesgo (8.7%) hasta cortar tangencialmente la frontera eficiente en el portafolio de Markowitz (máximo Sharpe). En este sentido, la teoría moderna de portafolios (MPT) dice que los inversores conservadores deben implementar una combinación del portafolio de máximo ratio de Sharpe con inversiones en el activo libre de riesgo, mientras que los inversores agresivos pueden apalancarse a la tasa libre de riesgo con el fin de invertir más del 100% en el portafolio de máximo ratio de Sharpe (Petzel, 2021).

Finalmente, el esquema de colores con base en el ratio de Sharpe evidencia la superioridad de los portafolios ubicados cerca a la parte superior de la frontera eficiente (colores verdes) respecto a aquellos ubicados en la parte baja de los portafolios factibles (colores amarillos y rojos).

Los resultados superiores del portafolio de Markowitz en términos del ratio de Sharpe confirma que la optimización media-varianza produce mejores resultados evaluados a través de esta métrica que la minimización aislada de la varianza. El ratio de Sharpe negativo del portafolio de mínima varianza puede atribuirse a las condiciones específicas del mercado entre 2019-2023, donde en términos generales las acciones listadas presentaron un comportamiento lateral, sin dejar de lado el impacto que tuvo la pandemia del COVID-19. Lo anterior sirve de evidencia para considerar la importancia de evaluar de forma conjunta tanto el riesgo como el retorno en la construcción de portafolios.

Antes de observar y analizar el desempeño de cada uno de los portafolios construidos a lo largo del año 2024, resulta conveniente tener en cuenta el desempeño de cada una de las acciones durante el año.

**Tabla 11. Rentabilidades Año 2024**

<b>Ticker</b>	<b>Rentabilidad Anual (2024)</b>
BCOLOMBIA	21,97%
BOGOTA	3,86%
BVC	-6,02%
CELSIA	29,34%
CEMARGOS	55,35%
CNEC	-61,78%
CONCONCRET	60,78%
CORFICOLCF	8,24%
ECOPETROL	-20,40%
ÉXITO	-48,50%
FABRICATO	-80,42%
GEB	34,28%
GRUBOLIVAR	11,04%
GRUPOARGOS	53,63%
GRUPOSURA	28,62%
ISA	13,63%
MINEROS	82,73%
NUTRESA	55,98%
PROMIGAS	51,86%
TERPEL	48,99%

Nota: Elaboración propia con base en datos de Yahoo Finance

La tabla 11 presenta las rentabilidades obtenidas por cada acción durante el año 2024. Los resultados evidencian un notable cambio respecto al comportamiento observado en el periodo de optimización 2019-2023 (Tabla 7), donde las rentabilidades anuales promedio oscilaron entre -16.86% (CNEC) y 17.22% (NUTRESA), mostrando una menor dispersión de rentabilidades.

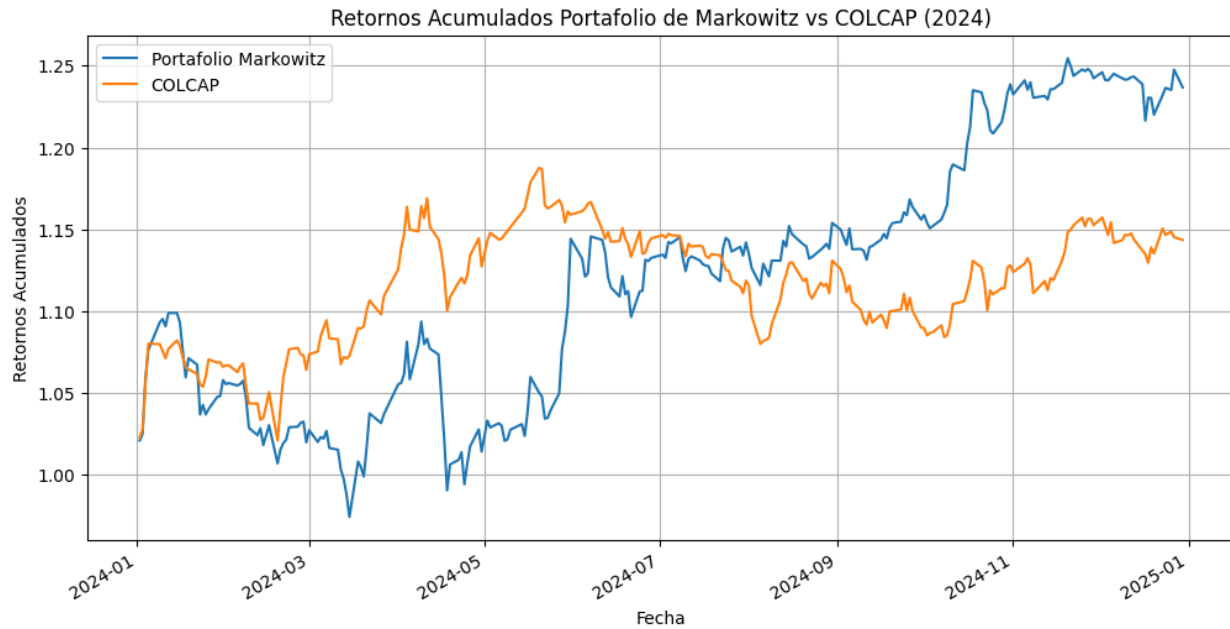
En contraste, el año 2024 registró una mayor variabilidad en las rentabilidades de las diferentes acciones, con valores que oscilan entre -80.42% (FABRICATO) y 82.73% (MINEROS). Sin embargo, es destacable que 15 de las 20 acciones consideradas tuvieron rentabilidades positivas, 11 de ellas con rentabilidades superiores al 20%, evidenciando así la dinámica favorable que experimentó el mercado bursátil colombiano durante 2024.

Dicho todo lo anterior, el siguiente paso será la optimización de los diferentes portafolios de inversión. Dada la información presente en la matriz de covarianza y en la rentabilidad esperada de cada una de las acciones, así como las restricciones de no utilizar apalancamiento, llevar a cabo únicamente operaciones de compra y una participación máxima de 20%, fueron obtenidos los siguientes resultados.

Por otra parte, resulta conveniente aclarar que, aunque durante el periodo 2019-2023 la tasa libre de riesgo (TES a 10 años) fue en promedio de 8.68%, durante el año 2024 esta experimentó un incremento considerable hasta ubicarse en una tasa promedio de 10.75%, siendo necesario utilizar esta cifra para calcular las métricas del portafolio y llevar a cabo los cálculos del modelo CAPM.

#### ***6.4.2 Portafolio Ratio de Sharpe***

El portafolio construido mediante la optimización de Markowitz enfocado en maximizar el ratio de Sharpe puede ser considerado como el punto de referencia para evaluar los demás portafolios. Tal como se aprecia en la figura 9, este portafolio logra superar durante los últimos meses del año la rentabilidad generada por el índice COLCAP.

**Figura 9. Desempeño Portafolio Ratio de Sharpe Comparado con Índice COLCAP - 2024**

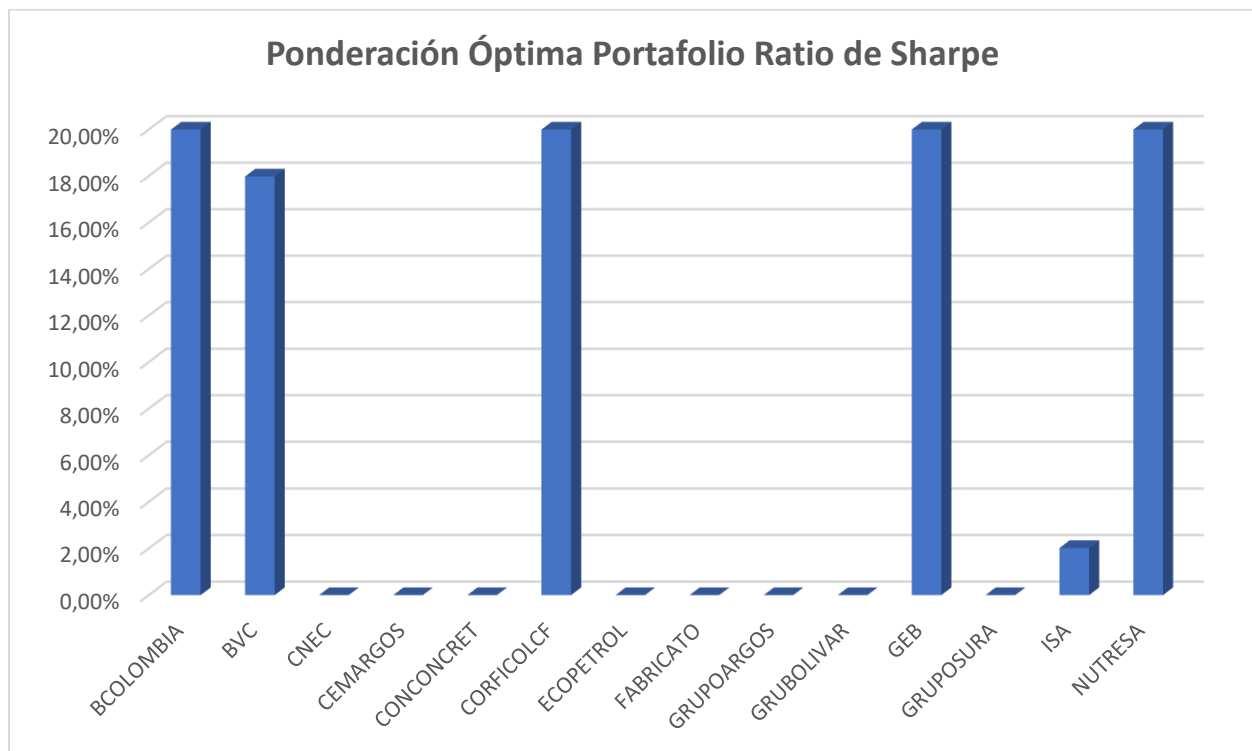
Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

Específicamente, el portafolio de Markowitz logra obtener una rentabilidad de 23.14%, superando por un margen de casi 10% el desempeño del COLCAP (14.11%), así como a los demás portafolios que serán evaluados posteriormente. Esta rentabilidad fue posible obtenerla manteniendo una volatilidad controlada de 16.37%, tan solo ligeramente superior al índice de referencia, demostrando así que la optimización media-varianza sigue siendo efectiva en el contexto del mercado accionario colombiano.

El ratio de Sharpe de 0.7568 representa un resultado positivo, siendo el mejor entre los tres portafolios tradicionales, esto gracias a su rentabilidad superior manteniendo controlado el nivel de riesgo, dejando claro que la optimización utilizada aún ofrece ventajas en la administración de portafolios. Por su parte, la beta de 0.7232 sugiere que la gestión activa del portafolio logra reducir ligeramente la exposición a los movimientos del mercado general.

En cuanto al alfa, cuyo valor fue del 9.96%, esto constituye evidencia clave de que aún en un mercado accionario como el colombiano, la teoría moderna de portafolios (MPT) cuenta con la capacidad de generar un valor agregado y rentabilidades superiores a aquellas esperadas por el modelo CAPM, considerando la exposición del portafolio al riesgo sistemático. De igual forma, tal resultado pasa a ser el objetivo que deberían superar los algoritmos de machine learning.

**Figura 10. Ponderación Óptima Portafolio Ratio de Sharpe**



Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República

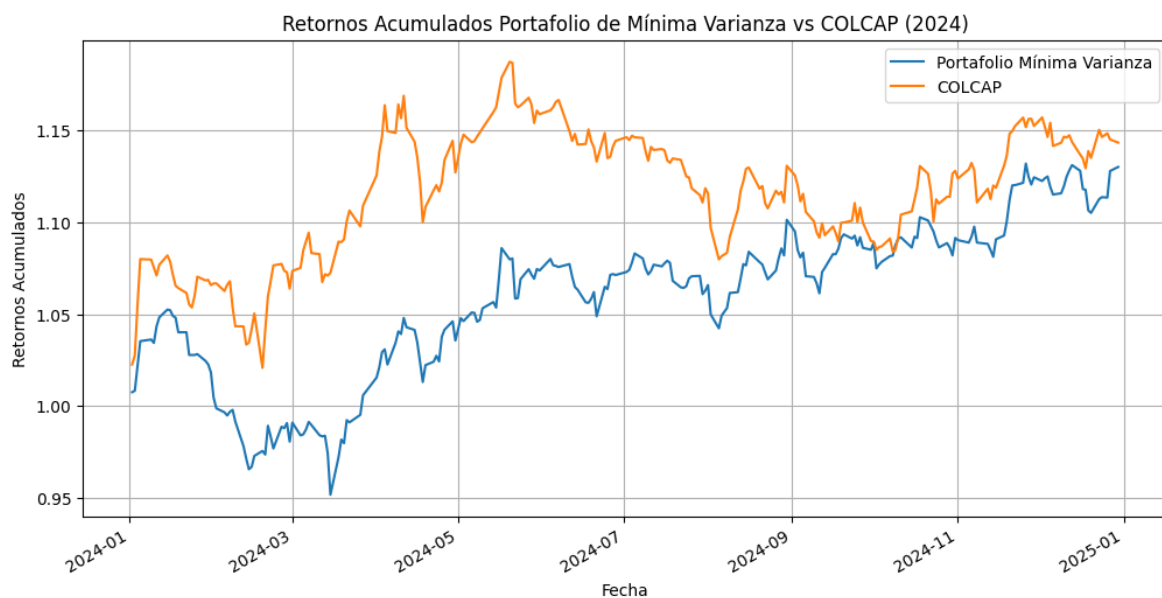
En cuanto a la composición del portafolio a lo largo del año 2024, esta es reflejada en la figura 10. Allí es posible destacar las participaciones con que cuentan las acciones de BCOLOMBIA, BVC, CORFICOLCF, GEB y NUTRESA, representando el 97.98% del portafolio (conviene recordar que la participación máxima que puede tener una acción es del 20%). Por su parte, la acción de ISA cuenta con el 2.02% del capital, mientras las demás cuentan con ponderaciones ligeramente superiores a 0%. Aunque esta asignación de capital concentrada en

únicamente cuatro acciones pueda parecer riesgosa, se debe tener en consideración que esta asignación está basada en el comportamiento histórico de las correlaciones y retornos de las diferentes acciones, así como la aplicación de la optimización media-varianza, garantizando una diversificación estratégica que logre maximizar la rentabilidad sin descuidar la volatilidad del portafolio.

### 6.4.3 Portafolio Mínima Varianza

El portafolio construido mediante la optimización media-varianza enfocado en minimizar la varianza representa una estrategia conservadora que busca priorizar la estabilidad sobre la rentabilidad, esto considerando además el hecho de que estimar los retornos esperados de cada acción es un procedimiento susceptible a errores de estimación que podrían afectar los resultados del portafolio (Swensen, 2009; Petzel, 2021). Dicho esto, la figura 11 presenta el comportamiento del portafolio de mínima varianza durante el año 2024, reflejando su comportamiento más estable respecto al índice COLCAP, aunque con una menor rentabilidad.

**Figura 11. Desempeño Portafolio Mínima Varianza Comparado con Índice COLCAP - 2024**

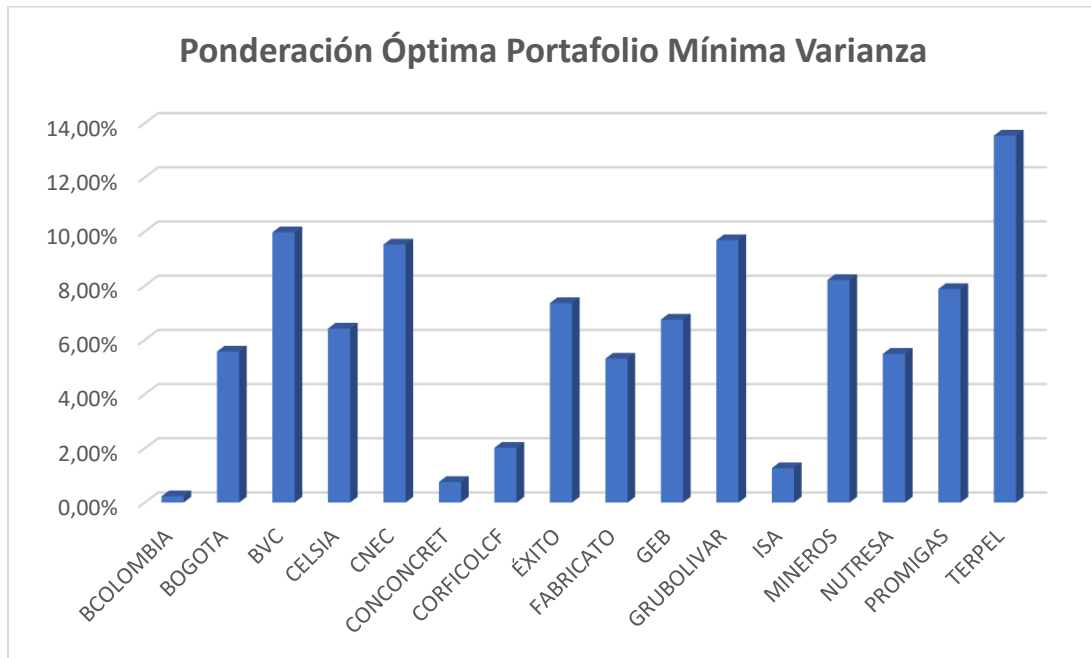


Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

Específicamente, el portafolio de mínima varianza logra obtener una rentabilidad de 13.04%, siendo ligeramente inferior a la experimentada por el índice COLCAP (14.11%), sin embargo, logrando reducir la volatilidad hasta el 10.50%, 2.73% menos que el COLCAP. Por tanto, el portafolio de mínima varianza cumple satisfactoriamente su objetivo de minimizar el riesgo, logrando a su vez mantener una rentabilidad atractiva y comparable a la ofrecida por el mercado.

Por su parte, el ratio de Sharpe de este portafolio logra ubicarse únicamente en 0.2188, cifra que permite evidenciar que este portafolio no genera una rentabilidad sobresaliente considerando su nivel de riesgo. Por el contrario, el ratio de Sharpe del índice COLCAP (0.2540) ofrece una mejor compensación al inversor manteniendo niveles similares de riesgo. En cuanto a la beta del portafolio esta equivale a 0.4139, consiguiendo la menor exposición de todos los portafolios a los movimientos del mercado general, resultado que está acorde con el objetivo de minimizar el riesgo.

El alfa de 0.90% constituye evidencia de que en el mercado bursátil colombiano, la estrategia de mínima varianza tiene una ligera capacidad de generar un valor agregado a la gestión de portafolios gracias a reducir de forma estratégica el nivel de riesgo respecto a los movimientos del mercado general (riesgo sistemático).

**Figura 12. Ponderación Óptima Portafolio Mínima Varianza**

Nota: Elaboración propia con base en datos de Yahoo Finance

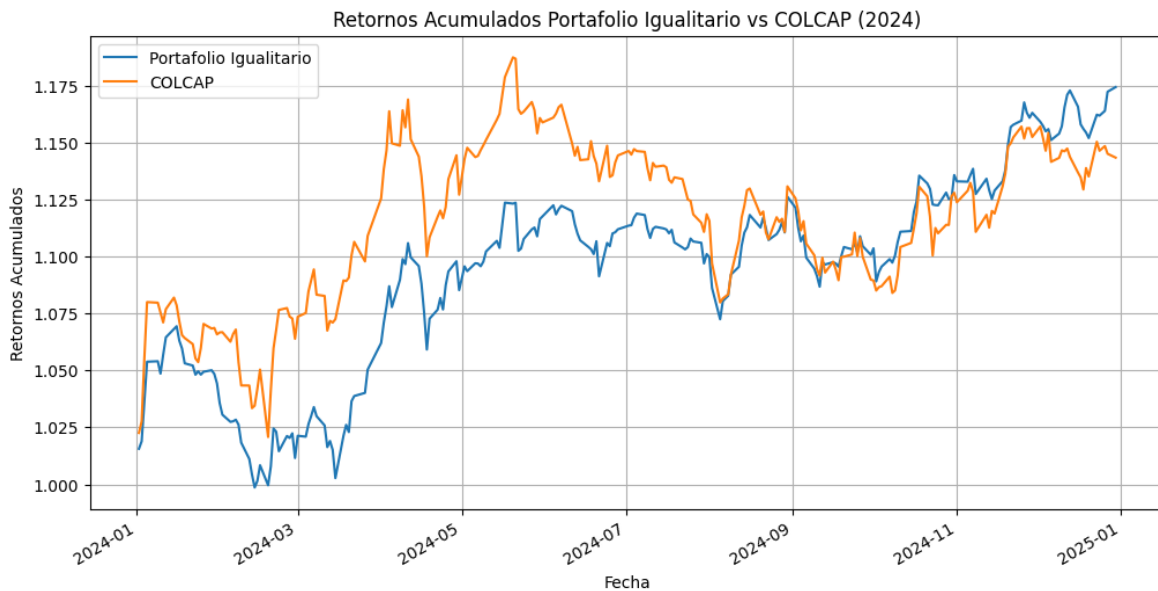
La composición del portafolio a lo largo del año 2024, reflejada en la figura 12, presentó una mayor diversificación que el portafolio de Markowitz. Específicamente, resaltan las acciones de TERPEL, BVC, GRUBOLIVAR, CNEC y MINEROS con participaciones mayores al 8%, mientras que las acciones de CORFICOLCF, ISA, CONCRET y BCOLOMBIA tienen participaciones menores al 2%.

### 6.4.3 Portafolio Ponderación Igualitaria

Finalmente, el portafolio de ponderación igualitaria no presenta ninguno de los potenciales errores de estimación presentes en la optimización media-varianza: los errores al estimar la matriz de covarianzas y los errores al estimar los retornos esperados. De esta forma, los portafolios de ponderación igualitaria tienden a presentar resultados más consistentes y estar exentos de errores (Swensen, 2009; DeMiguel et al., 2009; Petzel, 2021).

Dicho lo anterior, la figura 13 muestra el comportamiento del portafolio de ponderación igualitaria a lo largo del 2024, reflejando su menor volatilidad respecto al índice COLCAP, logrando aun así superar su rentabilidad durante las últimas semanas del año.

**Figura 13. Desempeño Portafolio Igualitario Comparado con Índice COLCAP - 2024**



Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

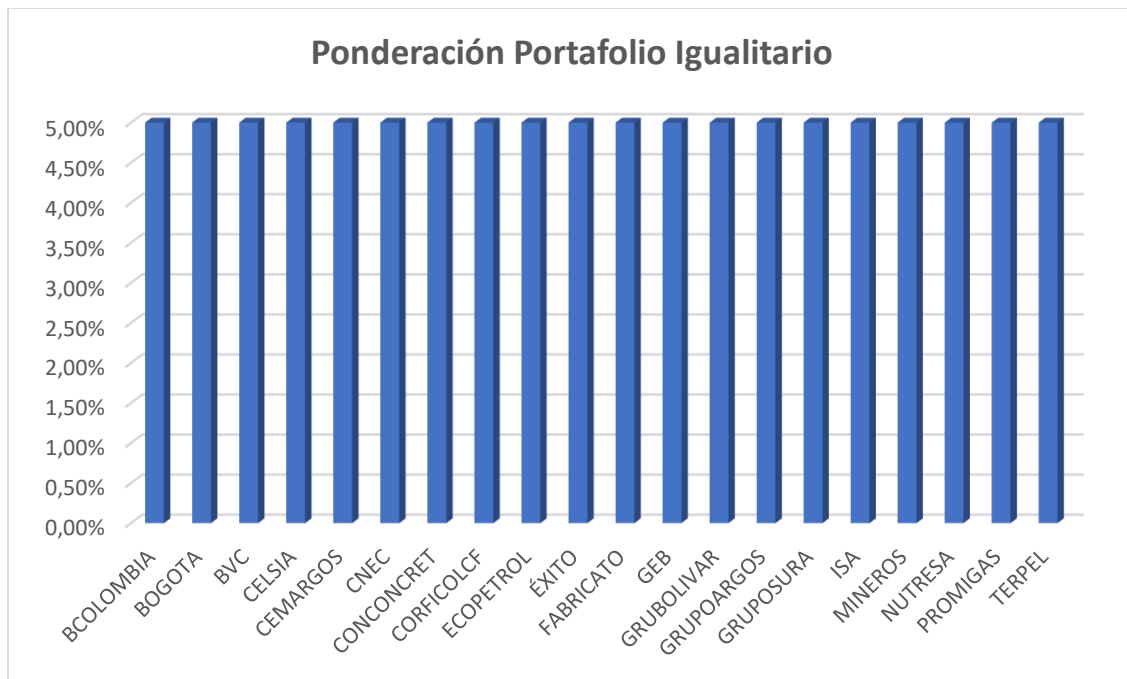
Específicamente, el portafolio de ponderación igualitaria logra alcanzar una rentabilidad de 17.16%, siendo esta 3.05% mayor a la experimentada por el índice COLCAP (14.11%). Sin embargo, es importante destacar el hecho de que la volatilidad logró alcanzar el 10.08%, una volatilidad no solo menor que la del índice de referencia (13.23%), sino además menor a la experimentada por el portafolio de mínima varianza (10.50%). Esto demuestra que, aunque la diversificación que divide el capital de forma igualitaria entre cada activo es una estrategia bastante simple, no significa que sus resultados sean deficientes, al contrario, al evitar posibles errores de estimaciones y no requerir de pronósticos de retornos puede resultar siendo una estrategia robusta.

El ratio de Sharpe de 0.6355 representa un resultado bueno, aunque no destacado, logrando ser superior al alcanzado por el COLCAP (0.2540) gracias al conseguir no solo aumentar la

rentabilidad sino también al reducir el riesgo, dejando claro que al evaluar el rendimiento ajustado al riesgo, estrategias simples de diversificación ofrecen ventajas significativas y fáciles de replicar. En cuanto a la beta de 0.578, esta sugiere que el portafolio tiende a experimentar fluctuaciones menores a las del mercado en general.

El alfa del portafolio, cuyo valor fue de 4.47% refleja la capacidad que tienen simples estrategias de diversificación para generar rendimientos adicionales por sobre aquellos esperados por el modelo CAPM dada la exposición al riesgo sistemático. De igual forma, esto permite considerar que estrategias básicas pueden llegar a competir con los resultados logrados por estrategias más complejas.

**Figura 14. Ponderación Portafolio Igualitario**



Nota: Elaboración propia con base en datos de Yahoo Finance

Finalmente, la composición del portafolio a lo largo del año 2024 está reflejada en la figura 14. Tal como su nombre lo indica, el portafolio de ponderación igualitaria asigna la misma ponderación a cada uno de los activos considerados en la muestra, siendo asignado el 5% a cada

uno de ellos dada la muestra de 20 acciones. Aunque esta asignación pueda parecer demasiado simple a primera vista, se debe tener presente que elimina totalmente los errores de estimación presentes en los métodos de optimización, sobre todo en mercados volátiles donde las correlaciones y rentabilidades de los activos fluctúan constantemente.

Una vez concluida la evaluación de portafolios construidos mediante la optimización media-varianza propuesta por Harry Markowitz y la estrategia de ponderación igualitaria, la siguiente sección se enfocará en el análisis de los portafolios construidos haciendo uso de algoritmos de machine learning. Posteriormente, todos los portafolios serán evaluados de manera comparativa con base en el alfa generado por cada uno, teniendo como referencia la fórmula planteada por el CAPM, al igual que teniendo en cuenta como métrica el ratio de Sharpe, con el fin de determinar cuál de las estrategias implementadas brinda los mejores resultados.

### **6.5 Portafolios Metodología Machine Learning**

Con el fin de garantizar la optimalidad de las series de tiempo de los retornos logarítmicos de cada una de las acciones consideradas, fue implementada la prueba Dickey-Fuller aumentada (ADF) para determinar que las series de tiempo cumplieran la condición de ser estacionarias, característica que facilita su modelación. Los resultados son presentados en la siguiente tabla:

**Tabla 14. P-Valores prueba ADF Serie de Tiempo Retornos Logarítmicos**

Acción	P Valor
BCOLOMBIA	0,00000
BOGOTA	0,00000
BVC	0,00000
CNEC	0,00000
CELSIA	0,00000
CEMARGOS	0,00000
CONCONCRET	0,00000
CORFICOLCF	0,00000
ECOPETROL	0,00000
FABRICATO	0,00000
GRUPOARGOS	0,00000
GRUBOLIVAR	0,00000
GEB	0,00000
ÉXITO	0,00000
GRUPOSURA	0,00000
ISA	0,00000
MINEROS	0,00000
NUTRESA	0,00000
PROMIGAS	0,00000
TERPEL	0,00000

Nota: Elaboración propia con base en datos de Yahoo Finance

Con base en la tabla 14, es posible confirmar que todas las series de tiempo son estacionarias, pudiendo continuar con el proceso de entrenamiento de modelos e implementación de la estrategia de inversión.

Una vez entrenados de acuerdo con los pasos establecidos en la sección de metodología, es evaluada la capacidad predictiva de cada uno de los algoritmos de machine learning, a través de las métricas MSE, RMSE y MAE promedio por algoritmo, pudiendo rápidamente comparar su precisión.

**Tabla 15. Métricas Promedio por Algoritmo**

Métricas Promedio por Algoritmo			
Algoritmo	MSE	RMSE	MAE
Random Forest	0,00058	0,0223	0,0158
XGBoost	0,00050	0,0208	0,0137
LSTM	0,00043	0,0201	0,0127

Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

De acuerdo con la tabla 15, existe una clara relación entre la complejidad de los algoritmos y su precisión predictiva. El algoritmo LSTM cuenta con el mejor desempeño general al conseguir un MSE promedio de 0.00043, RMSE de 0.0201 y MAE de 0.0127, siendo así el modelo más preciso al momento de predecir retornos logarítmicos. Por su parte, XGBoost presenta resultados ligeramente inferiores en términos de precisión al contar con un MSE de 0.00050, RMSE de 0.0208 y MAE de 0.0137. Finalmente, Random Forest muestra la menor precisión entre los tres algoritmos evaluados.

Estos resultados sirven de evidencia para considerar la idea de que, a mayor complejidad del algoritmo, mayor será la precisión en las predicciones generadas, lo cual potencialmente puede traducirse en mejores resultados por parte de los portafolios construidos a continuación.

Aunque las diferencias en las métricas de cada algoritmo puedan parecer mínimas (entre 0.31% y 0.10%) se debe tener presente que la estrategia de inversión toma decisiones todos los días de negociación, de forma que errores mínimos pueden llegar a tener un impacto significativo en los resultados finales de los portafolios. Un pequeño error de 1% o más en las predicciones puede marcar la diferencia entre comprar una acción que efectivamente tenga un comportamiento al alza o una que presente una caída en su precio, afectando la rentabilidad acumulada a lo largo del tiempo.

Sin embargo, antes de continuar a la siguiente sección resulta conveniente analizar con mayor detalle las métricas de precisión de cada algoritmo. Al considerar la información expuesta

de forma conjunta en las tablas 16, 17 y 18, es evidente que los resultados de los algoritmos no son homogéneos entre sí.

Inicialmente, es posible destacar la persistencia de errores elevados en la acción de FABRICATO en los tres algoritmos considerados, siendo algo menores en el caso de LSTM, de forma que el comportamiento de los retornos de esta acción requiera de modelos más complejos o implementar un nuevo entrenamiento con diferentes parámetros para generar predicciones más acertadas, esto a causa posiblemente de su menor liquidez, mayor volatilidad y dificultades financieras experimentadas por la empresa a lo largo del año.

De manera similar, la acción de ÉXITO muestra errores considerables en el caso particular del algoritmo RF, los cuales logran ser reducidos de forma exitosa por los algoritmos XGB y LSTM, reflejando así la existencia de posibles características intrínsecas de la acción que requieren de una mayor capacidad de aprendizaje. En las demás acciones, los tres algoritmos presentan resultados similares entre sí, manteniéndose la tendencia a presentar menores valores de MSE, RMSE y MAE a medida que los algoritmos utilizados cuentan con mayores capacidades. La heterogeneidad de resultados presente en los algoritmos, tal como ya fue mencionado, tiene potenciales implicaciones directas en el desempeño de los portafolios de inversión.

**Tabla 16. Métricas Algoritmo Random Forest**

<b>Algoritmo Random Forest</b>			
<b>Acción</b>	<b>MSE</b>	<b>RMSE</b>	<b>MAE</b>
BCOLOMBIA	0,000307	0,0175	0,0126
BOGOTA	0,000240	0,0155	0,0102
BVC	0,000282	0,0168	0,0075
CELSIA	0,000253	0,0159	0,0111
CEMARGOS	0,000331	0,0182	0,0129
CNEC	0,001035	0,0322	0,0234
CONCONCRET	0,000782	0,0280	0,0161
CORFICOLCF	0,000318	0,0178	0,0129
ECOPETROL	0,000300	0,0173	0,0128
ÉXITO	0,001468	0,0383	0,0339
FABRICATO	0,002648	0,0515	0,0429
GEB	0,000274	0,0166	0,0117
GRUBOLIVAR	0,000302	0,0174	0,0115
GRUPOARGOS	0,000222	0,0149	0,0109
GRUPOSURA	0,000570	0,0239	0,0163
ISA	0,000306	0,0175	0,0133
MINEROS	0,000478	0,0219	0,0151
NUTRESA	0,001014	0,0318	0,0178
PROMIGAS	0,000299	0,0173	0,0119
TERPEL	0,000261	0,0162	0,0107

Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

**Tabla 17. Métricas Algoritmo XGBoost**

Acción	Algoritmo XGBoost		
	MSE	RMSE	MAE
BCOLOMBIA	0,000300	0,0173	0,0122
BOGOTA	0,000234	0,0153	0,0100
BVC	0,000239	0,0154	0,0057
CELSIA	0,000254	0,0159	0,0109
CEMARGOS	0,000321	0,0179	0,0127
CNEC	0,000863	0,0294	0,0194
CONCRET	0,000804	0,0284	0,0158
CORFICOLCF	0,000328	0,0181	0,0129
ECOPETROL	0,000288	0,0170	0,0124
ÉXITO	0,000358	0,0189	0,0129
FABRICATO	0,002369	0,0487	0,0360
GEB	0,000257	0,0160	0,0112
GRUBOLIVAR	0,000300	0,0173	0,0113
GRUPOARGOS	0,000206	0,0143	0,0103
GRUPOSURA	0,000557	0,0236	0,0157
ISA	0,000293	0,0171	0,0130
MINEROS	0,000484	0,0220	0,0145
NUTRESA	0,000928	0,0305	0,0155
PROMIGAS	0,000319	0,0178	0,0107
TERPEL	0,000254	0,0159	0,0103

Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

**Tabla 18. Métricas Algoritmo LSTM**

Acción	Algoritmo LSTM		
	MSE	RMSE	MAE
BCOLOMBIA	0,000279	0,0167	0,0121
BOGOTA	0,000241	0,0155	0,0104
BVC	0,000252	0,0159	0,0079
CELSIA	0,000258	0,0161	0,0112
CEMARGOS	0,000417	0,0204	0,0152
CNEC	0,000868	0,0295	0,0194
CONCONCRET	0,000784	0,0280	0,0157
CORFICOLCF	0,000279	0,0167	0,0122
ECOPETROL	0,000331	0,0182	0,0139
ÉXITO	0,000347	0,0186	0,0122
FABRICATO	0,000643	0,0254	0,0043
GEB	0,000245	0,0156	0,0115
GRUBOLIVAR	0,000308	0,0176	0,0115
GRUPOARGOS	0,000234	0,0153	0,0110
GRUPOSURA	0,000573	0,0239	0,0161
ISA	0,000389	0,0197	0,0154
MINEROS	0,000543	0,0233	0,0164
NUTRESA	0,000960	0,0310	0,0169
PROMIGAS	0,000327	0,0181	0,0108
TERPEL	0,000247	0,0157	0,0102

Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

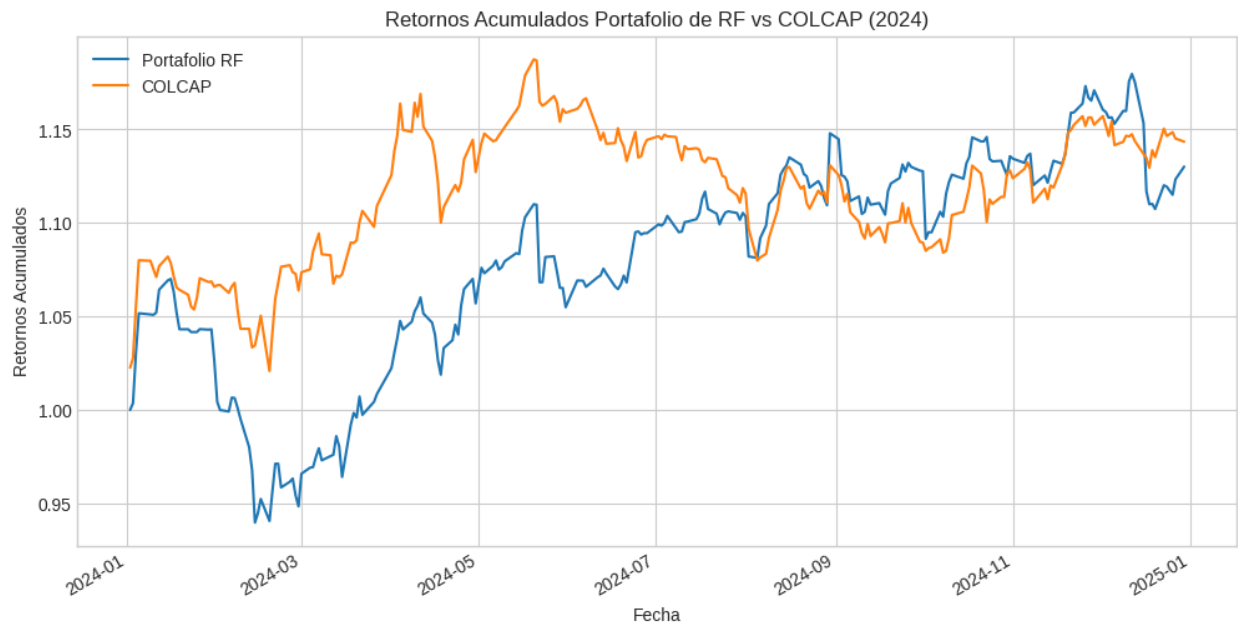
Con base en lo anterior, la siguiente sección evalúa como estas diferencias de precisión logran materializarse en términos de rentabilidad y riesgo en los portafolios construidos mediante cada algoritmo.

### 6.5.1 Portafolio Random Forest (RF)

El portafolio construido mediante el algoritmo Random Forest (RF) puede ser considerada la estrategia base entre los algoritmos de machine learning empleados, esto al ser uno de los más empleados de acuerdo con la revisión de literatura y su poca complejidad computacional, de forma que sirve como punto de referencia para evaluar algoritmos más complejos.

Tal como se aprecia en la figura 15, el comportamiento del portafolio RF no logra superar al índice COLCAP durante la mayor parte del año, aunque en el último trimestre este supera ligeramente la rentabilidad del índice, eventualmente antes de finalizar el año toma ciertas decisiones de inversión que afectan significativamente la rentabilidad acumulada.

**Figura 15. Desempeño Portafolio RF comparado con Índice COLCAP - 2024**



Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

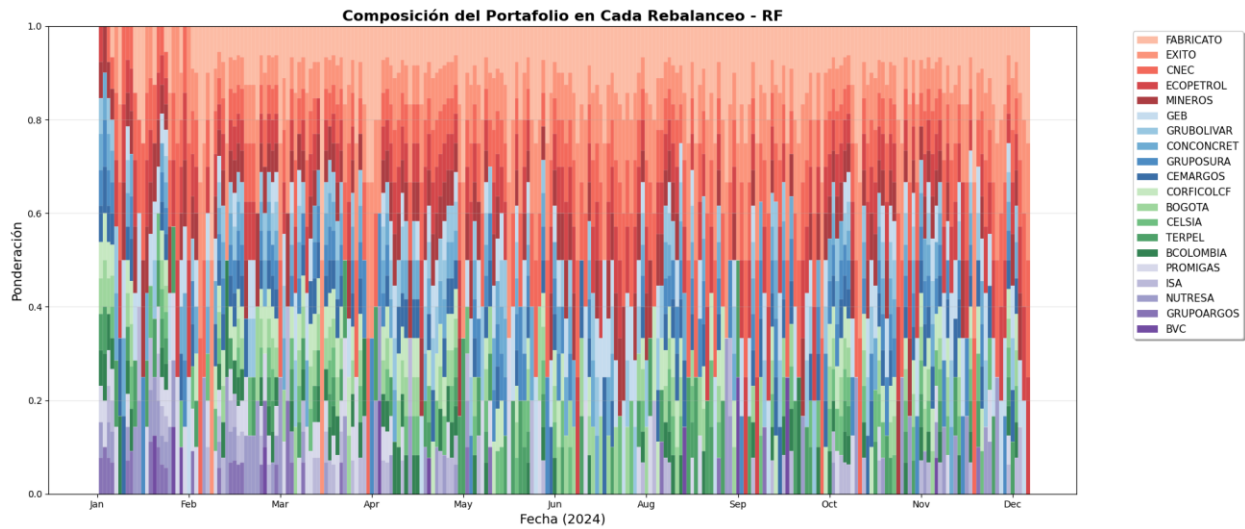
Dicho lo anterior, el portafolio RF generó un retorno de 12.80% en el año 2024, ligeramente inferior al 14.11% generado por el COLCAP. Este desempeño estuvo acompañado de una

volatilidad de 14.65%, superior en 1.42% al índice de referencia, reflejando así el hecho de que este portafolio presenta un desempeño ajustado al riesgo mucho menos eficiente que el mercado en general.

El ratio de Sharpe de 0.140 indica una rentabilidad ajustada al riesgo bastante baja, siendo superada tanto por el COLCAP (0.254), como por los algoritmos XGB (1.030) y LSTM (1.580). Por su parte, la beta de 0.631 sugiere una sensibilidad moderada a los movimientos del mercado en general, sin embargo, esta no llega a ser menor a la experimentada por los otros portafolios.

El alfa de -0.07% refleja el hecho de que el algoritmo RF no logra crear valor o generar una rentabilidad adicional por sobre aquella estimada por el modelo CAPM de acuerdo con la exposición al riesgo sistemático. Esto indica ciertas limitaciones en la capacidad del algoritmo RF para identificar adecuadamente oportunidades de inversión de corto plazo.

A pesar de que estos resultados no son los deseados considerando el objetivo de obtener rentabilidades superiores al mercado, si son útiles desde el punto de vista de demostrar que incluso algoritmos sencillos pueden generar resultados comparables a aquellos del mercado, abriendo paso a considerar la posibilidad de que algoritmos mejor optimizados o con mayores capacidades de aprendizaje obtengan los resultados deseados.

**Figura 16. Composición Portafolio RF - 2024**

Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

Finalmente, la figura 16 muestra la composición del portafolio a lo largo del año 2024. Al respecto, es posible mencionar que esta parece experimentar un comportamiento bastante dinámico sin patrones claros de inversión, sugiriendo así que el algoritmo responde a las fluctuaciones del mercado sin identificar tendencias consistentes.

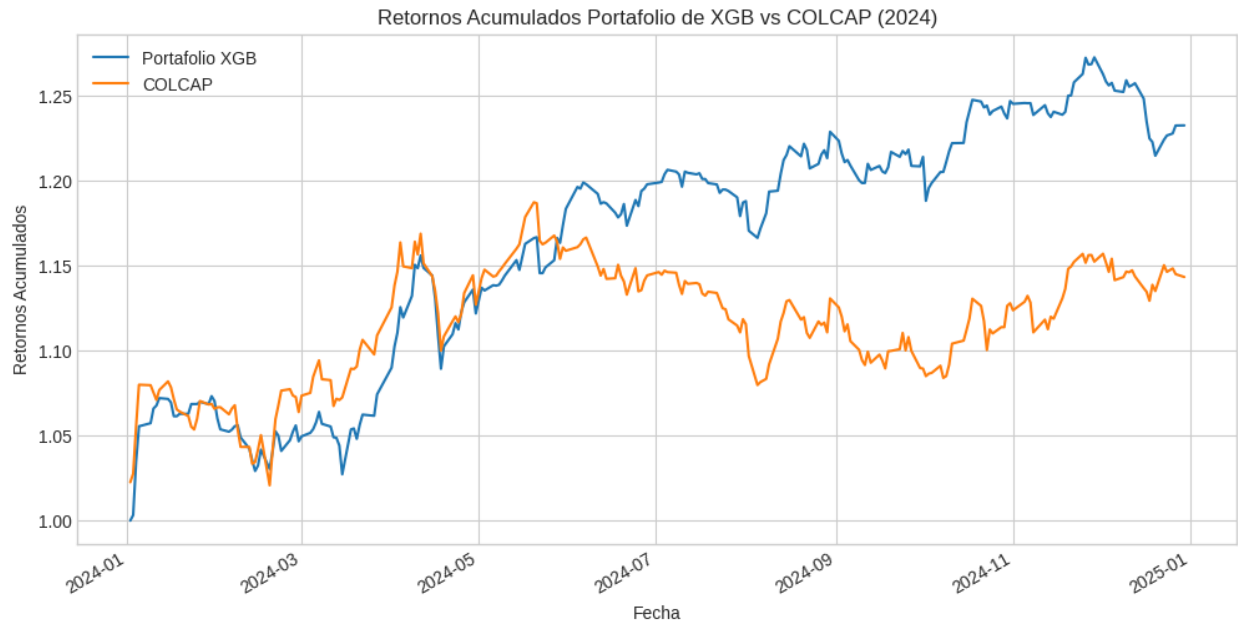
Sin embargo, resulta evidente la constante participación de las acciones de FABRICATO, ÉXITO, CNEC, ECOPETROL y MINEROS (color rojo), siendo las primeras tres de ellas acciones en las cuales el algoritmo RF presenta una menor precisión que las demás, por lo que la presencia de estas acciones y la baja precisión del modelo ayudan a explicar el rendimiento inferior que experimenta el portafolio RF, reflejando así el hecho de que errores en las predicciones se traducen en decisiones de inversión deficientes.

### 6.5.2 Portafolio XGBoost (XGB)

El portafolio construido mediante el algoritmo XGBoost presenta un desempeño considerablemente positivo en términos de rentabilidad durante el año 2024, ya sea en forma absoluta o ajustada al riesgo. Tal como se observa en la figura 17, durante la primera mitad del

año tanto el portafolio como el índice COLCAP presentan un desempeño similar, pero durante la segunda mitad del año el portafolio logra superar consistentemente al índice COLCAP.

**Figura 17. Desempeño Portafolio XGB comparado con Índice COLCAP - 2024**



Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

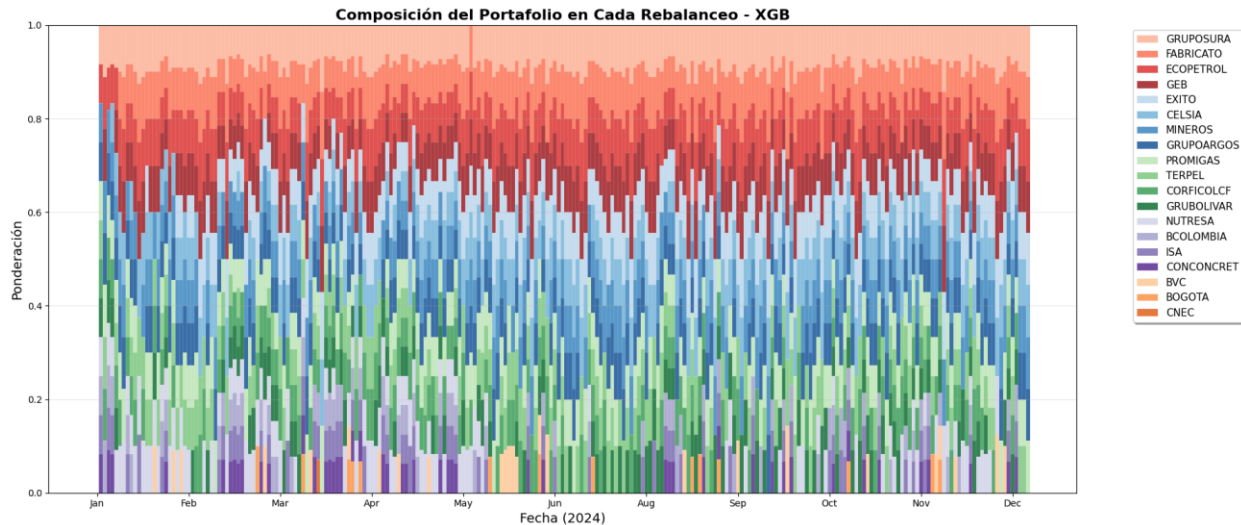
Específicamente, el portafolio XGB logra generar un retorno de 22.90%, significativamente superior al 14.11% del COLCAP. Esto a la vez que mantuvo una volatilidad o riesgo del 11.79%, 1.44% menor que el índice de referencia, de forma que se evidencia no solo una mejor rentabilidad sino también un menor nivel de riesgo.

Por su parte, el ratio de Sharpe de 1.030 indica que los inversionistas reciben más de una unidad de rentabilidad por cada unidad de riesgo asumido, superando al portafolio RF (0.140) y al índice COLCAP (0.254). De igual forma, la beta de 0.597 sugiere una menor sensibilidad a los movimientos del mercado, lo cual resulta deseable bajo condiciones de movimientos adversos.

Al respecto de alfa, su valor de 10.14% refleja que las predicciones generadas por el algoritmo XGB crea un valor agregado significativo más allá del esperado por la exposición al

riesgo sistemático, confirmando la capacidad del modelo y la estrategia para identificar y aprovechar oportunidades de inversión.

**Figura 18. Composición Portafolio XGB - 2024**



Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

La composición del portafolio XGB, presentado en la figura 18, exhibe un comportamiento dinámico, pero con mayor estructura respecto al portafolio RF. Específicamente, los rebalances parecen exhibir una respuesta más sistemática a las condiciones del mercado, dejando de lado el posible ruido existente, con una composición que tiende a ser relativamente constante durante gran parte del año.

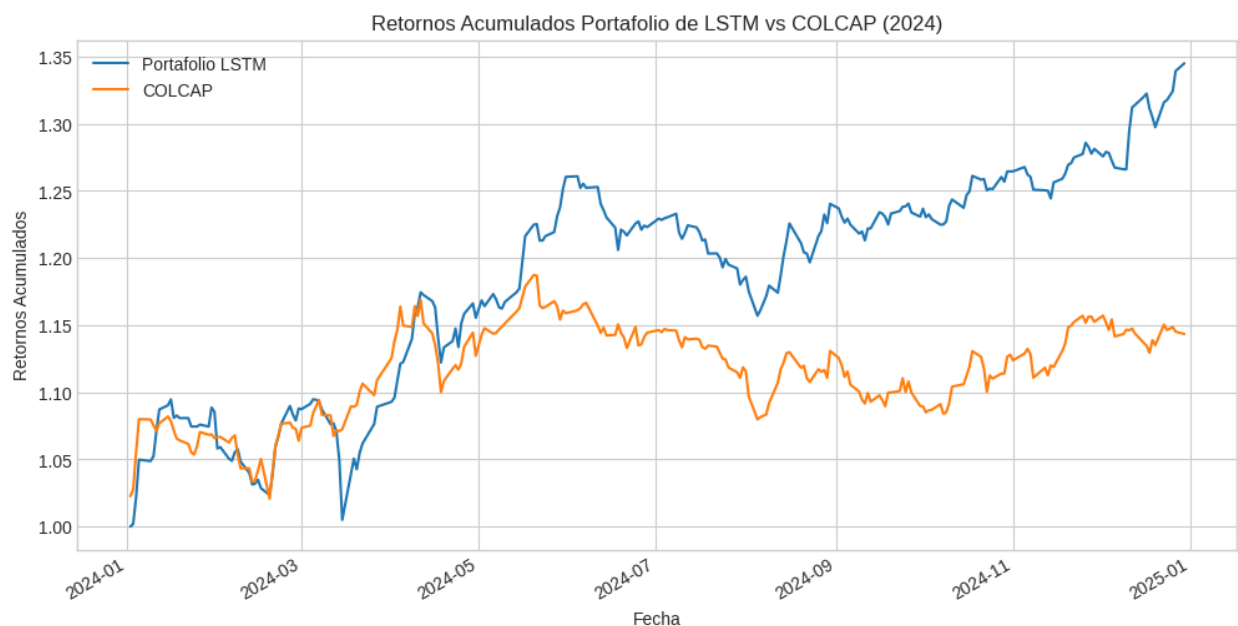
De igual forma, conviene resaltar la mayor diversificación que tiende a presentar el portafolio, en contraste con el portafolio RF, donde cierto grupo de acciones (FABRICATO, ÉXITO, CNEC, ECOPETROL y MINEROS) llegaban a tener una participación constante y elevada, en el caso del portafolio XGB las participaciones reflejan ser sostenidas durante un mayor periodo de tiempo, con bloques de participaciones constantes que pueden indicar mejores decisiones de inversión.

La menor dependencia de ciertas acciones, al igual que patrones de inversión más constantes a lo largo del tiempo pueden explicar parcialmente las mejoras en la rentabilidad generada por el portafolio XGB, a la vez que logra reducir parcialmente la exposición al riesgo.

### 6.5.3 Portafolio Long-Short Term Memory (LSTM)

El portafolio construido con base en el algoritmo LSTM cuenta con el desempeño más sobresaliente durante el año 2024 entre los tres portafolios construidos con algoritmos de ML. Tal como se evidencia en la figura 19, el portafolio LSTM muestra un comportamiento superior al índice COLCAP durante la mayor parte del año, destacando un rápido ascenso en su rentabilidad en el último trimestre.

**Figura 19. Desempeño Portafolio LSTM Comparado con Índice COLCAP - 2024**



Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

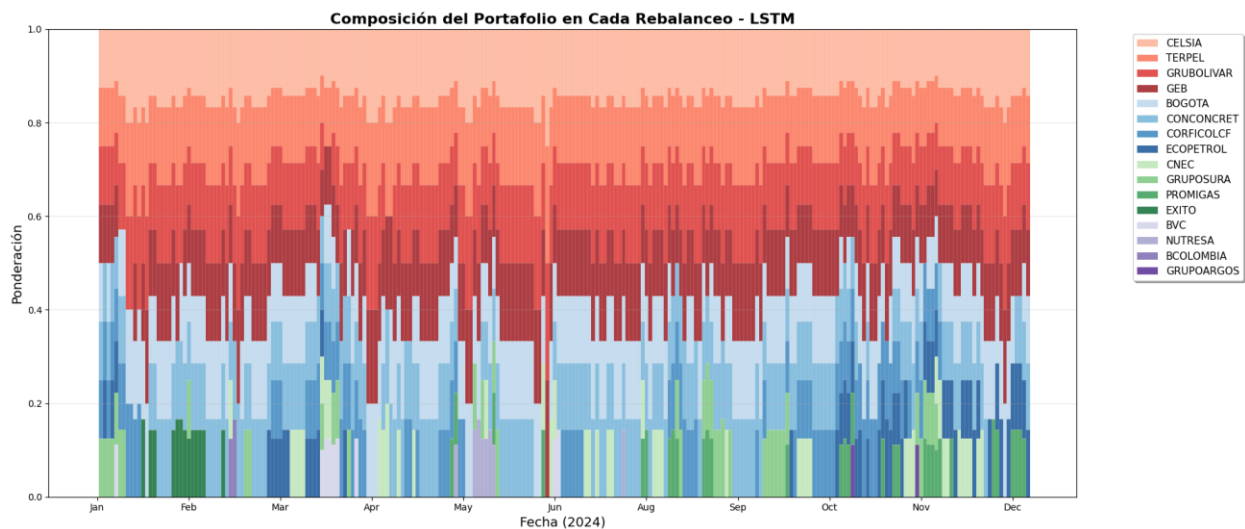
El portafolio LSTM logró obtener una rentabilidad de 33.94%, representando el mejor resultado entre las estrategias de ML y superando por un margen considerable al índice COLCAP (19.83%), así como al algoritmo XGB (11.04%) y RF (21.14%). Esta rentabilidad fue posible

obtenerla a la vez que mantuvo una volatilidad de 14.68%, tan solo 1.45% superior al índice de referencia, demostrando así una correcta gestión del riesgo dados los retornos generados.

Por su parte, el ratio de Sharpe de 1.5799 es el más alto entre los portafolios de ML, confirmando el hecho que LSTM ofrece la mejor rentabilidad ajustada al riesgo entre las diferentes alternativas evaluadas. Adicionalmente, la beta de 0.533 deja en evidencia la menor sensibilidad del portafolio respecto a los movimientos del mercado.

El alfa de 21.40% es notablemente superior a los demás, evidenciando así la capacidad de este algoritmo para captar patrones que permitan la creación de valor por parte de la gestión activa, más allá de lo esperado dada la exposición al mercado.

**Figura 20. Composición Portafolio LSTM - 2024**



Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

Partiendo de la información reflejada en la figura 20, la composición del portafolio LSTM se encuentra concentrada en un menor número de acciones que los portafolios RF y XGB. Específicamente, la composición muestra que el portafolio tiende a presentar composiciones sostenidas a lo largo del tiempo, sobre todo en las acciones de CELSIA, TERPEL, GRUBOLIVAR y GEB. Por su parte, las demás acciones presentan una participación que experimenta una mayor

fluctuación, siendo posible destacar las acciones de BOGOTA, CONCRET y CORFICOLCF.

La composición anterior podría ser un reflejo de la capacidad del algoritmo LSTM para identificar patrones de mayor duración en los retornos de las acciones que conforman la Bolsa de Valores de Colombia. Igualmente, esto permite que los rebalances en cada uno de los periodos de negociación tiendan a ser menos erráticos que en algoritmos como RF.

A pesar de que este portafolio tiene en promedio un menor número de acciones en su composición, por tanto, presentando así una mayor concentración, realiza un buen trabajo en mantener controlado el nivel de riesgo, ya que no solo cuenta con una beta de 0.533, sino que además su volatilidad a lo largo del año fue de 14.68%, ligeramente superior al índice de referencia.

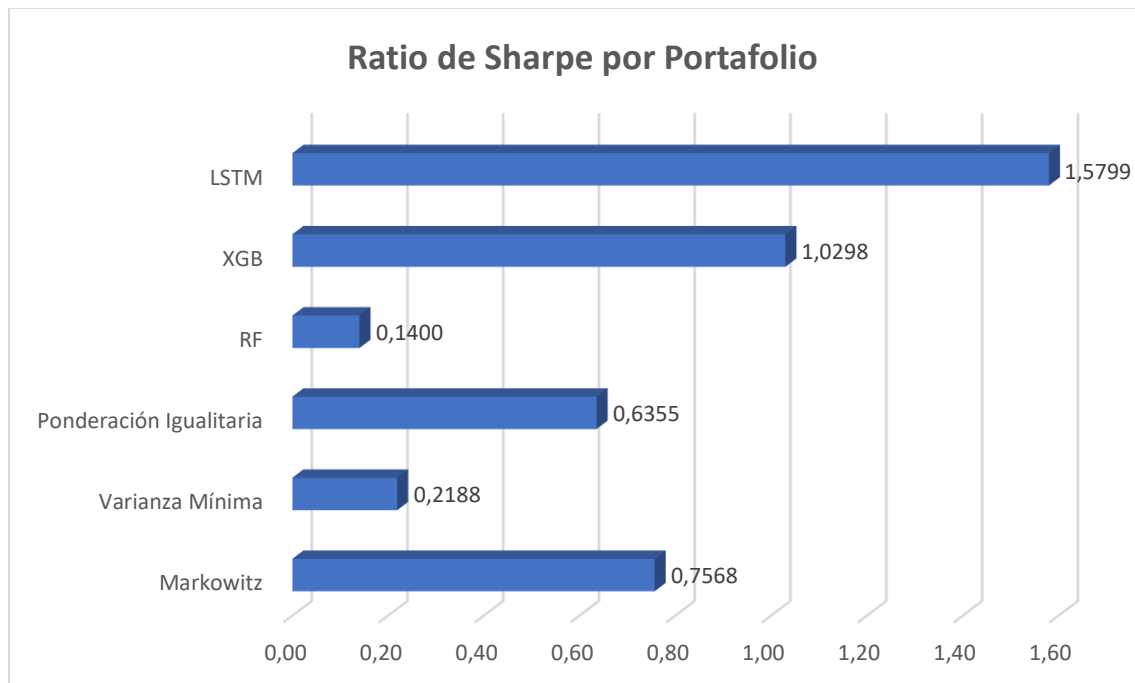
## 7. Discusión

Los resultados presentados en la sección anterior requieren de un mayor análisis y comparación para efectivamente comprender el significado de estos. Los resultados obtenidos se alinean consistentemente con los hallazgos reportados en la literatura internacional. Al igual que en los trabajos de López de Prado (2016), Zhang et al. (2021), Masuda (2024), Sutiene et al. (2024) y Gao et al (2024), los portafolios construidos con algoritmos de machine learning superan significativamente los resultados de aquellos portafolios basados en la optimización media-varianza, confirmando la robustez de estos enfoques en diferentes mercados y periodos de tiempo.

Teniendo como base la información presentada en la figura 21 y 22, el portafolio LSTM emerge como aquel con los mejores resultados, alcanzando un ratio de Sharpe de 1.5799 y un alfa

excepcional del 21.40%, superando incluso al portafolio de Markowitz (uno de los más implementados al momento de administrar portafolios), el cual cuenta con un ratio de Sharpe de 0.7568 y un alfa de 9.96%, resultados que a pesar de ser menores a los de LSTM siguen siendo positivos y valiosos desde el punto de vista de la teoría moderna de portafolios al lograr generar un valor adicional al inversor. Esta superioridad del 109% en el ratio de Sharpe está en línea con estudios previos: López de Prado (2016) obtuvo una mejora de 31.3% con la estrategia HRP respecto a MVO, Zhang et al. (2021) superaron en un 203% los resultados de los modelos tradicionales gracias al modelo E2E-LSTM, de forma que los resultados de la presente investigación se ubican en un punto medio.

**Figura 21. Comparación Ratio de Sharpe entre los diferentes Portafolios**

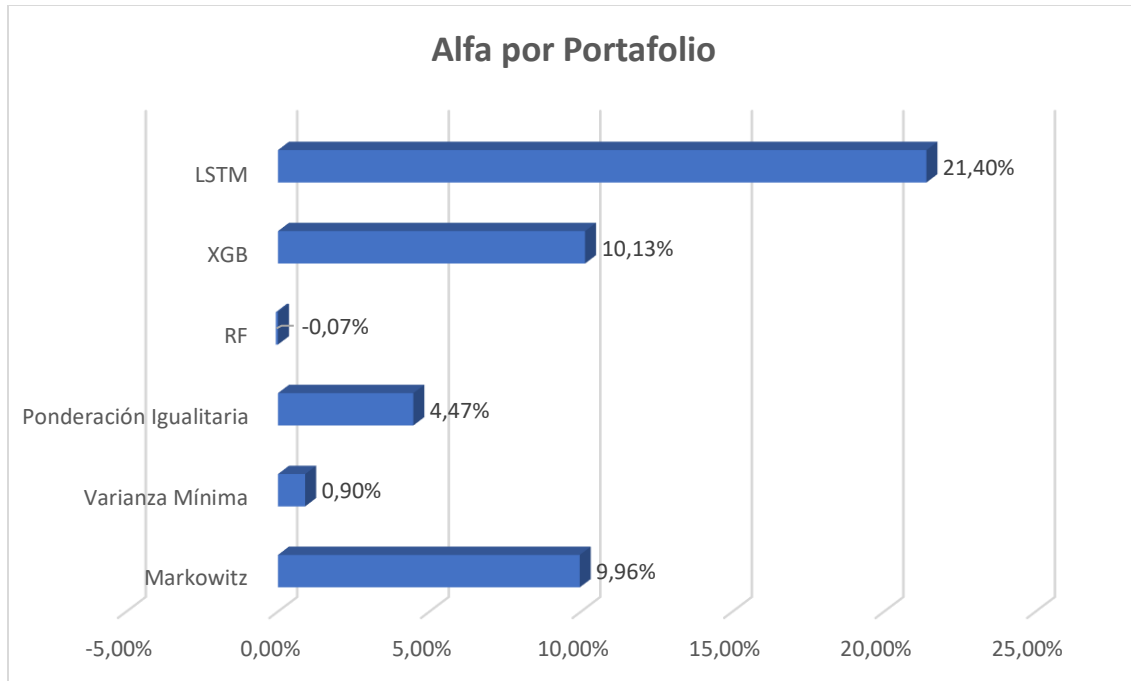


Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

Esta superioridad del LSTM es consistente con múltiples estudios académicos. Fischer y Krauss (2018) reportan un ratio de Sharpe extraordinario de 5.8 para LSTM en el índice S&P 500. Ghosh et al. (2022) confirmaron la superioridad de LSTM sobre Random Forest en el mismo

mercado (0.64% vs 0.54% de retornos diarios promedio). La jerarquía de resultados observada en la presente investigación coincide con la complejidad computacional de los diferentes algoritmos, respaldando los hallazgos de Masuda (2024) en el sentido de que modelos más complejos generaron alfas superiores.

**Figura 22. Comparación Alfa entre los diferentes Portafolios**



Nota: Elaboración propia con base en datos de Yahoo Finance y Banco de la República.

Juzgando los resultados, esto sugiere que algoritmos de machine learning sofisticados como LSTM pueden, al menos en las condiciones experimentadas por el mercado bursátil colombiano durante el 2024, superar el desempeño alcanzado por la tradicional optimización de portafolios.

Respecto a los resultados obtenidos por cada uno de los tres algoritmos empleados, resalta una clara mejora en el desempeño a medida que aumenta la complejidad computacional, específicamente: LSTM (Sharpe 1.5799 y alfa 21.40%); XGB (Sharpe 1.0298 y alfa 10.13%); RF (Sharpe 0.1400 y alfa -0.07%). De forma que de acuerdo con estos resultados sea posible

considerar la hipótesis de que la capacidad de los algoritmos para procesar información potencialmente se vea reflejada en la posibilidad de generar mejores resultados financieros.

Sin embargo, una mirada más detallada a estos resultados revela la presencia de rendimientos marginales decrecientes a medida que aumenta la complejidad computacional. Este patrón no ha sido documentado en la literatura previa, donde la mayor parte de los estudios se centran en demostrar la superioridad de los diferentes algoritmos en términos absolutos. La diferencia del rendimiento (evaluado a través del ratio de Sharpe) entre el portafolio XGB y RF (0.89) supera considerablemente a la diferencia existente entre el portafolio LSTM y XGB (0.55), indicando así la presencia de rendimientos marginales decrecientes. Dicho esto, al evaluar los rendimientos del portafolio y la complejidad del algoritmo utilizado, es posible sugerir la existencia de un punto de equilibrio, en el cual los resultados obtenidos resultan ser satisfactorios sin tener que enfrentar ciertas complicaciones técnicas, como lo son una mayor capacidad computacional y mayores tiempos de entrenamiento.

El desempeño destacado del portafolio LSTM puede atribuirse a la arquitectura especializada de este algoritmo, diseñado específicamente para trabajar con secuencias de datos, pudiendo así identificar más fácilmente ciertos patrones difíciles de capturar por algoritmos menos complejos.

Continuando con la presente discusión, resulta conveniente recalcar una vez más el hecho de que la estrategia utilizada en los portafolios de machine learning consiste en utilizar dichos algoritmos para generar las predicciones de retornos del siguiente día de negociación, asignando de forma igualitaria el capital entre aquellas con rendimientos positivos predichos, enfoque que difiere de aquel utilizado en las tradicionales estrategias de optimización.

Sin embargo, tal asignación de capital no es del todo una decisión infundada, puesto que en los resultados presentados se evidenció que el portafolio de ponderación igualitaria a pesar de su simplicidad obtuvo resultados positivos. Específicamente, esta estrategia, aunque simple, puede representar una ventaja considerable en aquellos mercados donde predecir la magnitud exacta del retorno resulta ser más difícil que predecir su dirección.

Aunque la información presentada hasta el momento refleja unos resultados positivos en términos de rentabilidad, es necesario juzgar que tan replicable es la estrategia de inversión implementada en condiciones reales del mercado. Inicialmente, el desempeño observado corresponde a un periodo de tiempo específico, año 2024, donde el mercado bursátil colombiano presentó un comportamiento destacado. Esta consideración está en línea con la realiza por Masuda (2024), quien señaló que el excepcional retorno del mercado americano en 2023 (24.14%) podría haber inflado los resultados. De forma similar, Fischer y Krauss (2018) experimentaron que LSTM generó rentabilidades solo hasta 2009, subrayando la importancia de validar los resultados en diferentes condiciones de mercado. Dicho esto, los resultados presentados puede que hubieran diferido considerablemente en caso de que el mercado haya presentado un comportamiento diferente, de modo que es necesario comprobar que desempeño podrían tener los diferentes algoritmos en otro entorno para poder garantizar su superioridad frente a las metodologías tradicionales de optimización.

Adicionalmente, conviene tener presente que comerciar en el mercado conlleva diferentes costos que socaban la rentabilidad de cualquier estrategia de inversión, aún más de aquellas estrategias activas. Específicamente, las comisiones de compra y venta, los spreads bid-ask y el impacto en el precio de cada orden (siendo estas dos últimas afectadas por la liquidez), son factores que cualquier estrategia de inversión efectiva y replicable en el mercado real deben tener en

consideración, sobre todo en mercados emergentes donde la poca liquidez puede convertirse en un gran obstáculo.

Profundizando un poco más en este aspecto, un factor crítico que limita la implementación práctica de estas estrategias en el mercado colombiano es la concentración de la liquidez en unas pocas acciones. El volumen de negociación se encuentra concentrado en un número reducido de acciones con alta capitalización bursátil, mientras que las demás acciones experimentan niveles de liquidez considerablemente menores, algunas llegando a registrar días sin transacciones.

Estas diferencias en la liquidez presentan desafíos importantes para las estrategias de gestión activa que requieren rebalanceos frecuentes. A medida que aumenta el tamaño del portafolio, surgen complicaciones cada vez mayores. Mientras las acciones de mayor liquidez pueden absorber órdenes considerables sin un impacto en sus precios, las acciones de menor liquidez enfrentan restricciones severas donde incluso portafolios de tamaño modesto puede llegar a representar una fracción significativa del volumen diario.

Esto significa que los algoritmos de ML al seleccionar acciones basándose únicamente en predicciones de retorno sin considerar las restricciones de liquidez, pueden estar inflando artificialmente los resultados obtenidos. Las estrategias que funcionan adecuadamente con portafolios pequeños pueden volverse imposibles de implementar a medida que el portafolio aumenta de valor, limitando severamente la aplicabilidad de los resultados obtenidos en condiciones reales de gestión de portafolios en el mercado local.

En relación con lo anterior, las estrategias que utilizan algoritmos de ML enfrentan mayores requerimientos desde un punto de vista técnico y computacional que no deben ser olvidados por el administrador de portafolios.

Antes de continuar la discusión resulta conveniente mencionar una importante limitación del trabajo desarrollado, siendo esta las limitaciones presentes en la fuente de datos utilizada, Yahoo Finance, utilizada ampliamente gracias a ser una herramienta gratuita. Sin embargo, la granularidad y la profundidad de los datos a los cuales es posible acceder a través de ella es limitada, siendo posible obtener únicamente información histórica con frecuencia diaria (precios de apertura, máximo, mínimo, cierre y volumen). La falta de acceso a datos intradía hace necesario realizar ciertos supuestos, como lo es asumir que todas las transacciones se ejecutan al precio de cierre, lo cual no es realista. Bajo condiciones reales, las acciones pueden llegar a presentar una volatilidad intradía significativa. Así mismo, ante la falta de información intradía es imposible implementar estrategias óptimas de ejecución de ordenes que busquen aprovechar periodos de mayor liquidez que brinden mejores condiciones a través de un spread bid-ask menor, así como garantizar la ejecución de ordenes de gran tamaño. Lo anterior va en línea con las recomendaciones de López de Prado (2018, 2018) sobre la importancia de utilizar datos de alta calidad.

Llevando la discusión al apartado de la gestión de activos (y por tanto estrategias de inversión activas), los hallazgos presentados en este trabajo tienen implicaciones prácticas para la gestión de activos en Colombia. La demostración de que el uso de algoritmos de machine learning tienen la capacidad para superar a las metodologías tradicionales, de forma que adoptar estos nuevos modelos y tecnología no solo pueden ayudar a generar una ventaja competitiva respecto a los demás administradores de portafolios, sino que también revela la necesidad de adoptar dichos algoritmos para mantener la relevancia en un entorno cada vez más competitivo. De forma que los resultados están alineados con el planteamiento de Petzel (2021) siendo una de las formas de mejorar los resultados (generar un alfa positivo) aquella de procesar los datos de una mejor forma, en este caso a través de modelos de machine learning. Esto a su vez refuerza la idea de que ante

un contexto donde la inversión pasiva gana fuerza rápidamente, la gestión activa sigue cumpliendo parte de su propuesta de valor.

Llevando la discusión a un apartado más técnico, resulta conveniente reconocer las limitaciones implícitas en la metodología utilizada en el proceso de construcción y optimización de los modelos de machine learning. La tarea por realizar, entrenar tres algoritmos diferentes para cada una de las veinte acciones consideradas en la muestra, impone restricciones prácticas sobre el ajuste y perfeccionamiento de cada modelo.

De cara a investigaciones futuras es indispensable abordar las limitaciones expuestas a lo largo del documento, siendo estas la inclusión de los costos de transacción y liquidez, al igual que la validación de los algoritmos en diferentes periodos de tiempo y condiciones de mercado. Así mismo, estas direcciones de investigación están alineadas con tendencias identificadas en la literatura: Implementar modelos híbridos siguiendo a Zhang et al. (2021) y Masuda (2024), validar en múltiples periodos de tiempo abordando las limitaciones identificadas por Fischer y Krauss (2018), e incorporar datos alternativos aprovechando las capacidades resaltadas por López de Prado (2018), Bartram et al. (2020), y Gao et al. (2024).

## **8. Conclusiones**

La presente investigación logró evaluar comparativamente el desempeño de portafolios de inversión construidos mediante algoritmos de machine learning versus metodologías tradicionales de optimización en el mercado accionario colombiano durante el periodo 2024. A través de la implementación de tres algoritmos de machine learning (Random Forest, XGBoost y LSTM) y tres portafolios tradicionales (Markowitz, mínima varianza y ponderación igualitaria), los resultados demuestran de manera concluyente que los algoritmos de machine learning,

particularmente LSTM, superan significativamente las metodologías tradicionales tanto en términos de rentabilidad absoluta como de rentabilidad ajustada al riesgo.

El portafolio LSTM alcanzó una rentabilidad del 33.94% con un ratio de Sharpe de 1.5799, superando al mejor portafolio tradicional (Markowitz: 23.14%, Sharpe 0.7568) por un margen del 46.7% en la rentabilidad y del 108.7% en el ratio de Sharpe. Estos resultados confirman la hipótesis inicial sobre la superioridad de los enfoques de machine learning en el contexto específico del mercado de valores colombiano.

La revisión de literatura reciente sobre la aplicación de ML en la gestión de portafolios identificó tres áreas principales de aplicación: procesamiento de datos alternativos, generación de predicciones y optimización de asignación de capital. Esta revisión evidenció la limitada investigación en mercados emergentes como el colombiano, justificando la relevancia del presente estudio.

La implementación de los algoritmos de machine learning para la predicción de retornos diarios de 20 acciones del índice COLCAP mostró una clara jerarquía en la precisión: LSTM (MAE: 0.0127); XGBoost (MAE: 0.0137); Random Forest (MAE: 0.0158), estableciendo la base para el desempeño obtenido por los portafolios. Esta jerarquía confirma la relación existente entre complejidad computacional y capacidad predictiva evidenciada en la literatura (Bartram et al., 2021; Gao et al., 2024; Kumari, 2024; Sutiene et al., 2024).

Los seis portafolios construidos y evaluados mediante backtesting con datos del 2024 permitieron una comparación directa entre los enfoques tradicionales y los algoritmos de ML, proporcionando evidencia empírica sobre su desempeño relativo. El análisis utilizando el ratio de Sharpe y alfa de Jensen como métricas principales reveló que dos algoritmos de ML generaron alfas positivos (LSTM: 21.40%, XGBoost: 10.13%), mientras que Random Forest presentó un alfa

negativo marginal (-0.07%). Es importante destacar que el portafolio de Markowitz también logró un alfa de 9.96%, reflejando que las metodologías tradicionales aún mantienen capacidad para generar valor adicional en el mercado de valores colombiano.

Respecto a las demás investigaciones consultadas, el presente trabajo identifica la presencia de rendimientos marginales decrecientes en función de la complejidad computacional y tiempo de entrenamiento. Mientras la diferencia en el ratio de Sharpe entre XGBoost y Random Forest fue de 0.89 puntos, la diferencia entre LSTM y XGBoost fue de solo 0.55 puntos. Lo anterior puede llevar a considerar la existencia de un punto de equilibrio entre la complejidad computacional (por tanto, en el tiempo de entrenamiento) y las mejoras de rentabilidad.

Esta evidencia tiene implicaciones prácticas importantes. Para aquellos inversores o gestores de portafolios con capacidades computacionales limitadas, XGBoost puede brindar un balance óptimo entre desempeño y complejidad computacional, mientras que inversores o gestores con mayores capacidades computacionales pueden maximizar sus resultados implementando un algoritmo LSTM.

La metodología implementada, que combina predicciones de algoritmos de ML con asignación igualitaria de capital entre activos con retornos positivos predichos, demuestra que enfoques relativamente simples pueden generar resultados capaces de competir con metodologías más complejas evidenciadas en la literatura. A diferencia de metodologías como el “End-to-End” propuesto por Zhang et al. (2021) que optimizan directamente las ponderaciones del portafolio, esta estrategia evita las mayores complejidades técnicas y computacionales de tal optimización mientras mantiene resultados superiores. Así mismo, el rebalanceo diario permite al portafolio adaptarse continuamente a las condiciones del mercado, característica particularmente valiosa en mercado volátiles.

A pesar de estos resultados prometedores, es importante reconocer las limitaciones del estudio. Este análisis se limita al año 2024, periodo caracterizado por un desempeño bastante positivo del mercado de valores colombiano. Para llegar a generalizar estos resultados se requiere su validación en diferentes condiciones de mercado, incluyendo por ejemplo periodos de alta volatilidad, tendencias bajistas e incluso de crisis financieras. Los hallazgos de Fischer y Krauss (2018), donde LSTM perdió su efectividad después de 2009, subraya la importancia de llevar a cabo esta validación.

Las restricciones impuestas por Yahoo Finance, especialmente la ausencia de datos intradía, limitan la implementación de mejores y más realistas estrategias de ejecución que tengan en consideración la liquidez y los costos de transacción reales. En este sentido, el supuesto de ejecución al precio de cierre, aunque necesario para el presente trabajo, representa una limitación que debe abordarse en investigaciones futuras con acceso a datos de mayor granularidad.

De igual forma, los algoritmos más complejos requieren de capacidades computacionales y conocimientos técnicos avanzados que pueden no estar disponibles para todos los inversores o gestores de portafolios, limitando parcialmente el acceso a estos enfoques en el mercado colombiano.

Investigaciones futuras deben priorizar la validación de los algoritmos en múltiples periodos de tiempo y condiciones de mercado, siguiendo las recomendaciones de las observaciones realizadas por Fischer y Krauss (2018) y Masuda (2024). La implementación de modelos híbridos que combinen fortalezas de diferentes algoritmos representa una dirección con posibles resultados positivos que permitan optimizar el trade-off entre complejidad y desempeño.

La incorporación de datos alternativos aprovechando las capacidades identificadas en la literatura (Simonian et al., 2018; Lopez de Prado, 2019; Bartram et al., 2021; Gao et al., 2024;

Kumari, 2024) podría mejorar significativamente las capacidades predictivas de los algoritmos. Así mismo, el análisis detallado de costos de transacción reales y su incorporación en la optimización de los portafolios representa una línea de investigación crucial para la implementación práctica de estas estrategias.

De igual forma, la extensión del análisis a otros mercados emergentes latinoamericanos permitiría validar la generalización de resultados e identificar características regionales específicas que faciliten o limiten la implementación de estrategias de ML en la gestión de portafolios.

En conclusión, esta investigación demuestra de manera concluyente que los algoritmos de machine learning, particularmente LSTM, pueden generar valor significativo en la gestión de portafolios en mercados emergentes como Colombia. Los resultados van más allá de la expectativa inicial de documentar la superioridad del ML sobre los métodos tradicionales de optimización, ya que también fue posible identificar la presencia de rendimientos marginales decrecientes, cuya consideración puede ser de especial utilidad para inversores con capacidades computacionales limitadas. La adopción de algoritmos de machine learning en la gestión de portafolios emerge no solo como una innovación tecnológica, sino como una evolución necesaria para mantener la competitividad y generar valor consistentemente para los inversionistas.

**Referencias Bibliográficas**

- Ahmed, S., Alshater, M. M., Ammari, A. E., & Hammami, H. (2022). Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*, 61, 101646. <https://doi.org/10.1016/j.ribaf.2022.101646>
- Argos. (2024). *Sobre Nosotros*. <https://argos.co/contexto/>
- Bahrammirzaee, A. (2010). A comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert system and hybrid intelligent systems. *Neural Computing and Applications*, 19(8), 1165-1195. <https://doi.org/10.1007/s00521-010-0362-z>
- Banco de Bogotá. (2013). *Informe de Gestión—Primer Semestre 2013*.
- Bartram, S. M., Branke, J., & Motahari, M. (2020). *Artificial intelligence in asset management*. CFA Research Institute Foundation.
- Bartram, S. M., Branke, J., Rossi, G. D., & Motahari, M. (2021). *MACHINE LEARNING FOR ACTIVE PORTFOLIO MANAGEMENT*.
- Bernstein, P. L. (1999). Why The Efficient Market Offers Hope To Active Management. *Journal of Applied Corporate Finance*, 12(2), 129-136.
- BVC. (2025). *Nuestra Empresa*. <https://bvc.com.co//nuestra-empresa>
- Canacol Energy Ltd. (2025). *Quienes Somos*. Canacol Energy Ltd. <https://canacolenergy.com/es/about-us/who-we-are/>
- Celsia. (2024). *Qué hacemos*. Celsia. <https://www.celsia.com/es/quienes-somos/que-hacemos/>
- Concreto. (2024). *Nosotros*. *Concreto*. <https://concreto.com/nosotros/>

Corficolombiana. (2025). *About us | Corfi*. <https://www.corfi.com/documents/20123/0/corfi-corporate-presentation-2023.pdf>

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *The Review of Financial Studies*, 22(5), 1915-1953. <https://doi.org/10.1093/rfs/hhm075>

Ecopetrol. (2021, febrero 11). *Acerca de Ecopetrol*. <https://www.ecopetrol.com.co/wps/portal/Home/es/NuestraEmpresa/QuienesSomos/acerca-de-ecopetrol>

Fabricato. (2024). Fabricato | Pasión y compromiso en la industria textil colombiana. *Fabricato*. <https://fabricato.com/nuestra-empresa/>

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25(2), 383-417. <https://doi.org/10.2307/2325486>

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3-56. [https://doi.org/10.1016/0304-405X\(93\)90023-5](https://doi.org/10.1016/0304-405X(93)90023-5)

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1-22. <https://doi.org/10.1016/j.jfineco.2014.10.010>

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654-669. <https://doi.org/10.1016/j.ejor.2017.11.054>

- Freitas, F. D., De Souza, A. F., & De Almeida, A. R. (2009). Prediction-based portfolio optimization model using neural networks. *Neurocomputing*, 72(10-12), 2155-2170. <https://doi.org/10.1016/j.neucom.2008.08.019>
- Gao, H., Kou, G., Liang, H., Zhang, H., Chao, X., Li, C.-C., & Dong, Y. (2024). Machine learning in business and finance: A literature review and research opportunities. *Financial Innovation*, 10(1), 86. <https://doi.org/10.1186/s40854-024-00629-z>
- GEB. (2025). *Conoce GEB*. Grupo Energía Bogotá. <https://www.grupoenergiabogota.com/conoce-geb#Talento-GEB>
- Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*.
- Ghosh, P., Neufeld, A., & Sahoo, J. K. (2022). Forecasting directional movements of stock prices for intraday trading using LSTM and random forests. *Finance Research Letters*, 46, 102280. <https://doi.org/10.1016/j.frl.2021.102280>
- Grupo Argos. (2024). *Sobre Grupo Argos*. <https://www.grupoargos.com/quienes-somos/sobre-grupo-argos/>
- Grupo Bancolombia. (2025). *About us*. <https://www.grupobancolombia.com/corporate/about-us>
- Grupo Bolívar. (2012). *¿Quiénes Somos?* <https://www.grupobolivar.com.co/>
- Grupo Éxito. (2025). *Quiénes somos*. <https://www.grupoexito.com.co/es/quienes-somos>
- Grupo Nutresa. (2022). *Acerca de nosotros—Nutresa*. <https://gruponutresa.com/quienes-somos/acerca-de-nosotros/>

Grupo Sura. (2024, diciembre 17). *Sobre Grupo SURA*. Grupo SURA.

<https://www.gruposura.com/nuestra-compania/sobre-grupo-sura/>

Gumelar, A. B., Setyorini, H., Adi, D. P., Nilowardono, S., Latipah, Widodo, A., Wibowo, A. T.,

Sulistiyono, M. T., & Christine, E. (2020). Boosting the Accuracy of Stock Market

Prediction using XGBoost and Long Short-Term Memory. *2020 International Seminar on*

*Application for Technology of Information and Communication (iSemantic)*, 609-613.

<https://doi.org/10.1109/iSemantic50169.2020.9234256>

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios.

*Applied Stochastic Models in Business and Industry*, 33(1), 3-12.

<https://doi.org/10.1002/asmb.2209>

Heaton, J. B., Polson, N., & Witte, J. (2016). *Deep Learning for Finance: Deep Portfolios* (SSRN

Scholarly Paper 2838013). <https://doi.org/10.2139/ssrn.2838013>

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *SCIENCE*,

349(6245), 261-266. <https://doi.org/10.1126/science.aaa8685>

ISA. (2025). ISA. ISA. <https://isa.co/es/grupo-isa/>

Isichenko, M. (2021). *Quantitative Portfolio Management: The Art and Science of Statistical*

*Arbitrage*.

Jang, J., & Seong, N. (2023). Deep reinforcement learning for stock portfolio optimization by

connecting with modern portfolio theory. *Expert Systems with Applications*, 218,

119556. <https://doi.org/10.1016/j.eswa.2023.119556>

Jensen, M. (1967). *The Performance of Mutual Funds in the Period 1945-1964* (SSRN Scholarly

Paper 244153). Social Science Research Network. <https://doi.org/10.2139/ssrn.244153>

- Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2), 689-702. <https://doi.org/10.1016/j.ejor.2016.10.031>
- Kumari, S. (2024). AI-Enhanced Portfolio Management: Leveraging Machine Learning for Optimized Investment Strategies in 2024. *Journal of Informatics Education and Research*, 4(3), Article 3. <https://doi.org/10.52783/jier.v4i3.1487>
- Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics*, 47(1), 13-37. <https://doi.org/10.2307/1924119>
- Lo, A. W. (2004). The Adaptive Markets Hypothesis. *The Journal of Portfolio Management*, 30(5), 15-29. <https://doi.org/10.3905/jpm.2004.442611>
- Lopez de Prado, M. (2016). *Building Diversified Portfolios that Outperform Out-of-Sample* (SSRN Scholarly Paper 2708678). <https://doi.org/10.2139/ssrn.2708678>
- López de Prado, M. (2018). *Advances in Financial Machine Learning*.
- Lopez de Prado, M. (2019). *Beyond Econometrics: A Roadmap Towards Financial Machine Learning* (SSRN Scholarly Paper 3365282). <https://doi.org/10.2139/ssrn.3365282>
- Malkiel, B. G. (2020). *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing*.
- Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*, 7(1), 77-91. <https://doi.org/10.2307/2975974>

Masuda, J. S. (2024). *Portfolio Optimization Using a Hybrid Machine Learning Stock Selection Model* [Thesis, Massachusetts Institute of Technology].

<https://dspace.mit.edu/handle/1721.1/157186>

Michaud, R. O. (1989). *The Markowitz Optimization Enigma: Is «Optimized» Optimal?* (SSRN Scholarly Paper 2387669). Social Science Research Network.

<https://doi.org/10.2139/ssrn.2387669>

Nazareth, N., & Reddy, Y. V. R. (2023). Financial applications of machine learning: A literature review. *EXPERT SYSTEMS WITH APPLICATIONS*, 219, 119640.

<https://doi.org/10.1016/j.eswa.2023.119640>

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., Reuter, U., Gama, J., & Gandomi, A. H. (2020). Data Science in Economics: Comprehensive Review of Advanced Machine Learning and Deep Learning Methods. *Mathematics*, 8(10), Article 10.

<https://doi.org/10.3390/math8101799>

Osorio, M., & Rico, J. (2024). *Optimización de un portafolio de inversión con acciones del COLCAP aplicando técnicas de machine learning*.

<https://repository.eafit.edu.co/entities/publication/fc9f5d5d-d76e-45bd-8f50-67a996a208cf>

Petzl, T. E. (2021). *Modern Portfolio Management: Moving Beyond Modern Portfolio Theory*.

Promigas. (2024). *Sobre Promigas*.

[https://www.promigas.com/Paginas/Nuestra\\_Empresa/ESP/SobrePromigas.aspx](https://www.promigas.com/Paginas/Nuestra_Empresa/ESP/SobrePromigas.aspx)

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3), 341-360. [https://doi.org/10.1016/0022-0531\(76\)90046-6](https://doi.org/10.1016/0022-0531(76)90046-6)

- Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90, 106181. <https://doi.org/10.1016/j.asoc.2020.106181>
- Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk\*. *The Journal of Finance*, 19(3), 425-442. <https://doi.org/10.1111/j.1540-6261.1964.tb02865.x>
- Simonian, J., Lopez de Prado, M., & Fabozzi, F. J. (2018). *Order from Chaos: How Data Science is Revolutionizing Investment Practice* (SSRN Scholarly Paper 3246407). <https://papers.ssrn.com/abstract=3246407>
- Starmer, J. (2022). *The StatQuest Illustrated Guide To Machine Learning*.
- Sutiene, K., Schwendner, P., Sipos, C., Lorenzo, L., Mirchev, M., Lameski, P., Kabasinskas, A., Tidjani, C., Ozturkkal, B., & Cerneviene, J. (2024). Enhancing portfolio management using artificial intelligence: Literature review. *Frontiers in Artificial Intelligence*, 7. <https://doi.org/10.3389/frai.2024.1371502>
- Swensen, D. F. (2009). *Pioneering Portfolio Management: An Unconventional Approach to Institutional Investment, Fully Revised and Updated*.
- Terpel. (2025). *¿Quiénes somos?* Terpel. <https://www.terpel.com/quienes-somos>
- Treynor, J. L. (1962). *Jack Treynor's «Toward a Theory of Market Value of Risky Assets»* (SSRN Scholarly Paper 628187). <https://doi.org/10.2139/ssrn.628187>
- Zhang, C., Zhang, Z., Cucuringu, M., & Zohren, S. (2021). *A Universal End-to-End Approach to Portfolio Optimization via Deep Learning* (arXiv:2111.09170). arXiv. <https://doi.org/10.48550/arXiv.2111.09170>

Zhang, Y. (2024). Stock Price Prediction Method Based on XGboost Algorithm. En *Proceedings of the 2022 International Conference on Bigdata Blockchain and Economy Managemen.*

[https://doi.org/10.2991/978-94-6463-030-5\\_60](https://doi.org/10.2991/978-94-6463-030-5_60)