APLICACIONES DE MACHINE LEARNING PARA LA PREDICCIÓN DE RENDIMIENTOS Y GESTIÓN DE UN PORTAFOLIO DE ACTIVOS FINANCIEROS EN LA BVC

Juan Pablo Gómez Bravo
Lizeth Fernanda Suarez Mensa
Katherin Juliana Quiñones Losada

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

David Arango Londoño

Director

Jurado

Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, Julio 07/2023

Autor:

Juan Pablo Gomez Bravo
Lizeth Fernanda Suarez Mensa
Katherin Juliana Quiñones Losada

Título del Trabajo de Grado: APLICACIONES DE MACHINE LEARNING PARA LA PREDICCIÓN DE RENDIMIENTOS Y GESTIÓN DE UN PORTAFOLIO DE ACTIVOS FINANCIEROS EN LA BVC.

Director:

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

David Arango Londoño

Firma del Director del Trabajo de Grado

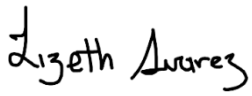
Santiago de Cali, 29 de Mayo de 2023

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

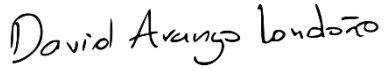
Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado "Aplicaciones de Machine Learning para la predicción de rendimientos y gestión de un portafolio de activos financieros en la BVC", el cual será realizado por el (la) estudiante Lizeth Suarez con código 00020314366, Juan Pablo Gómez, con código 0008972429 y Katherin Quiñones con código 00020353189 perteneciente al énfasis en N/A, bajo la dirección del profesor David Arango Londoño.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,



Lizeth F. Suarez
C.C. 1005829261 de Cali,Valle



David Arango Londoño
C.C. 1130586950 de Cali, Valle

Juan Pablo Suarez
C.C. 1032402668 de Bogotá,Cundinamarca

Katherin J. Quiñones
C.C. 1143864388 de Cali, Valle

APLICACIONES DE MACHINE LEARNING PARA LA PREDICCIÓN DE RENDIMIENTOS Y GESTIÓN DE UN PORTAFOLIO DE ACTIVOS FINANCIEROS EN LA BVC.

Lizeth Fernanda Suarez Mensa
Código 00020314366
Katherin Juliana Quiñones Losada
Código 00020353189
Juan Pablo Gómez Bravo
Código 0008972429

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director(a)
David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JULIO 01 DE 2023

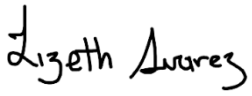
Santiago de Cali, 29 de Mayo de 2023

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado "Aplicaciones de Machine Learning para la predicción de rendimientos y gestión de un portafolio de activos financieros en la BVC", el cual será realizado por el (la) estudiante Lizeth Suarez con código 00020314366, Juan Pablo Gómez, con código 0008972429 y Katherin Quiñones con código 00020353189 perteneciente al énfasis en N/A, bajo la dirección del profesor David Arango Londoño.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,



Lizeth F. Suarez
C.C. 1005829261 de Cali, Valle

David Arango Londoño
C.C. 1130586950 de Cali, Valle

Juan Pablo Suarez
C.C. 1032402668 de Bogotá

Katherin J. Quiñones
C.C. 1143864388 de Cali, Valle

RESUMEN

La teoría de portafolio se desarrolla en un ambiente de incertidumbre, es decir los inversionistas no conocen con certeza el resultado que puede obtenerse al realizar una inversión diversificada, por ende, este trabajo busca gestionar el riesgo de pérdida, mitigando la posibilidad de obtener rendimientos inferiores a los esperados, por lo anterior se realizó la estimación de volatilidad de cada uno de los activos que componen COLCAP, cotizados en la Bolsa de Valores de Colombia (BVC), mediante varias metodologías tradicionales y de Machine Learning buscando el mejor desempeño de dicha estimación, una vez obtenidos los resultados más eficientes por cada activo, se realizan múltiples comparaciones para determinar las covarianzas de los portafolios factibles.

Para dar solución a la problemática, se hizo uso de un área fundamental del Machine Learning (Aprendizaje Automático) como lo es la predicción de series temporales, ya que contiene componentes de volatilidad, tendencia, etc. intrínsecas a su comportamiento. Una serie de observaciones tomadas cronológicamente en el tiempo se conoce como Serie de Tiempo. Se trabajarán con series de tiempo financieras a partir de las acciones más importantes de la BVC, con el objetivo de optimizar portafolios de inversión, que puedan resultar como solución a grandes inconvenientes que involucran cuantificar, medir e inclusive controlar el riesgo ante la exposición de inversión.

Una vez obtenidos los resultados de pronóstico, mediante el Modelo de Markowitz, (independientemente de los activos seleccionados), se obtiene la distribución ideal de la inversión o distribución eficiente bajo mínimo riesgo de pérdida. Adicionalmente el inversionista cuenta con la posibilidad de calcular una frontera eficiente, fruto de asignar rendimientos mayores que el rendimiento obtenido para el portafolio de mínimo riesgo, así como la información de su Rendimiento, Varianza y VaR del portafolio.

Finalmente, mediante Power BI, el inversionista podrá contar con el comportamiento de las tendencias de volatilidad de los activos de renta variable analizados en este trabajo, la posibilidad de elegir hasta un máximo de 5 activos de su interés y obtener la distribución eficiente de su inversión, así como la información referente a su portafolio mencionada previamente.

INDICE DE FIGURAS

Figura 1 Diagrama de una celda LSTM [15] Fuente: Elaboración Propia 22
Figura 2 Mockup Tablero Sección: Análisis Descriptivo 34
Figura 3 Mockup Tablero Sección: Selección Portafolio 35
Figura 4 Proceso Metodológico, Fuente: Elaboración propia 36
Figura 5 Separación de la data set en entrenamiento y test para cada serie de las acciones de la BVC 37
Figura 6 Gráfico Descriptivo: Precio vs Rendimiento (Primeras Acciones) 39
Figura 7 Gráfico Descriptivo: Precio vs Rendimiento (Complemento) 40
Figura 8 División de la data set en entrenamiento y prueba para cada conjunto de datos 42
Figura 9 Comparativo serie de prueba Vs Predicción del modelo GARCH 44
Figura 10 Comparativa proyecciones del SVR vs datos de validación 46
Figura 11 Comparativa proyecciones del XGBoost vs datos de validación y pronóstico (10 acciones) 48
Figura 12 Portafolio optimo conformado por 5 acciones 52
Figura 13 Pagina 1 Dashboard: Análisis descriptivo de la serie 53
Figura 14 Pagina 2 Dashboard: Definición de la rentabilidad esperada del portafolio escogido por el inversor 54
Figura 15 Pagina 3 Dashboard: Definición del portafolio optimo dado un tamaño de este 54

INDICE DE TABLAS

Tabla 1 Relación Objetivo, Tarea Desarrollada y Resultado 29
Tabla 2 Descripción de series de las acciones consideradas 29
Tabla 3 Tabla: Variación diaria precios de cierre Activo EXITO. 30
Tabla 4 Descriptivas Precio Cierre 37
Tabla 5 Descriptivas logaritmo de la variacion del precio cierre 38
Tabla 6 Hiperparametros y métricas del modelo base 43
Tabla 7 Hiperparámetros y resultados del ajuste del SVR 45
Tabla 8 Hiperparámetros y resultados del ajuste del XGboost 47
Tabla 9 Hiperparametros y resultados del ajuste de Redes Neuronales (LSTM) 49
Tabla 10 Modelo escogido para cada serie 50
Tabla 11 Series con la menor varianza en el periodo estudiado 52

TABLA DE CONTENIDO

INTRODUCCIÓN 8
1. DEFINICIÓN DEL PROBLEMA 9
1.1. PLANTEAMIENTO DEL PROBLEMA 9
1.2. FORMULACIÓN DEL PROBLEMA 10
2. OBJETIVOS DEL PROYECTO 11
2.1 OBJETIVO GENERAL 11
2.2 OBJETIVOS ESPECÍFICOS 11
3. MARCO TEORICO Y ANTECEDENTES 12
3.1. MARCO TEÓRICO 12
3.1.1. Marco Teórico Contextual 12
3.1.2. Marco Teórico Aplicado 14
3.1.2.1. Modelos 15
3.1.2.2. Gestión de Portafolios de inversión 22
3.1.2.3. Métricas de evaluación de los modelos: 23
3.2. ANTECEDENTES 24
4. METODOLOGÍA 28
4.1. Desarrollo de la metodología 28
4.2. Consolidación de la información 29
4.3. Preparación de los datos. 30
4.4. Proceso de modelación. 31
 Selección del modelo con mejor ajuste 32
4.5. Definición del portafolio de inversión 33
5. RESULTADOS 36
5.1 Definición del conjunto de entrenamiento y validación 36
5.2 Estimación de los distintos modelos de Machine Learning 41
5.2.1 Estimación del modelo base 41
5.2.2 Ajuste del modelo mediante Support Vector Regression 45
5.2.3 Ajuste del modelo mediante XGboost 47
5.2.4 Ajuste del modelo mediante Redes Neuronales 49
5.3 Selección del mejor modelo para cada serie 50

5.4 Definición de portafolio de inversión 52
5.5 Construcción del Dashboard en Power BI 53
6. CONCLUSIONES Y TRABAJOS FUTUROS 55
6.1. CONCLUSIONES 55
6.2. TRABAJOS FUTUROS 55
7. REFERENCIAS BIBLIOGRÁFICAS 60

INTRODUCCIÓN

En la actualidad, la gestión de portafolios de inversión se ha convertido en una necesidad para los individuos que buscan manejar su capital de manera eficiente. Una de las alternativas más comunes es la inversión en portafolios de activos de renta variable, sin embargo, la conformación de estos puede resultar un reto debido a la gran cantidad de opciones disponibles y a la dependencia del inversionista y su actitud frente al riesgo. La diversificación de portafolios es una de las estrategias para alcanzar carteras rentables y generadoras de altos retornos, pero siempre se debe tener en cuenta la gestión del riesgo para lograr las rentabilidades esperadas y minimizar el riesgo de pérdida.

La generación de portafolios ha dado origen a diferentes modelos que buscan optimizar el conjunto de títulos de valor disponibles, considerando criterios de decisión como el riesgo asociado y la rentabilidad esperada. Aunque el enfoque clásico de estos modelos ha cubierto algunas fases del proceso, la computación financiera ha surgido como una nueva técnica para procesar grandes volúmenes de información y dar respuesta a la necesidad de los inversionistas.

El Machine Learning se presenta como una herramienta que permite realizar predicciones en escenarios donde el enfoque clásico se queda corto, generando valor para los inversionistas al prever un portafolio diversificado que estabilice la pérdida y minimice el riesgo observado, sin afectar la rentabilidad esperada.

Este trabajo de grado tiene como objetivo evaluar diferentes metodologías de Machine Learning para la estimación de volatilidad de activos financieros y la optimización de la inversión en los mismos, con el fin de generar un portafolio de inversión eficiente y factible en la bolsa de valores de Colombia, que genere una rentabilidad esperada mínima con el menor riesgo posible en una ventana de tiempo determinada, esto a partir de la estimación de volatilidades mediante métodos tradicionales y de aprendizaje de máquina, logrando resultados más precisos y confiables y con la posibilidad de iterar un volumen mayor de información, finalmente el inversionista obtiene la distribución eficiente de su inversión en la selección de un portafolio factible de hasta 5 activos y obteniendo información relevante de su portafolio tales como Rendimiento, Varianza y VaR.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

El mercado financiero a nivel mundial se ha visto impactado en varias ocasiones, llegando a generar grandes pérdidas en el campo de inversión. Por ejemplo, la crisis financiera mundial que ocurrió en el 2008, ocasionada inicialmente por el colapso de Lehman-Brothers, trajo consigo un impacto a gran escala sobre los mercados financieros, afectando tanto a economías desarrolladas como a economías de menor nivel, generando una recesión. No era algo que tan siquiera se hubiese visualizado, probablemente, podía parecer como una idea fuera de la realidad, pero sucedió y demostró que, a nivel mundial, no había una preparación ante un problema de este calibre.

Bajo este contexto inicial, el mercado financiero adopta una disciplina como la ciencia de datos, que desde la extracción de valor de los datos y la formulación de preguntas o hipótesis, soportados en métodos mucho más robustos y potentes que la modelación tradicional, logrando capacidades extraordinarias como el manejo de grandes volúmenes de datos, conectividad entre diferentes fuentes de datos y puntos de interacción con los datos, permitiendo a los usuarios finales una mejor capacidad de toma de decisiones frente a escenarios de alta incertidumbre.

Teniendo en cuenta la problemática anteriormente planteada respecto a cómo las metodologías usadas para el análisis del mercado financiero o en la construcción de un portafolio de inversión, se pueden quedar cortas a la hora de prever inestabilidad en estos mercados y cómo estos posiblemente se pueden ver solventados haciendo uso de técnicas sofisticadas de Machine Learning (ML), es de vital importancia conocer algunos conceptos y algunas problemáticas, que intervienen en la medición del comportamiento de la volatilidad los activos financieros.

En finanzas, la volatilidad de los activos financieros es un problema fundamental a solucionar, ya que el estudio y anticipación de esta, permite a los inversionistas una mejor aproximación a la estimación de sus rentabilidades futuras y por ende una reducción al riesgo de pérdida, que sin duda es un objetivo principal en un mercado dinámico y de alta variabilidad. La volatilidad se mide mediante la desviación estándar de los rendimientos del activo y es el indicador del riesgo de una inversión en un activo financiero, por lo cual se hizo necesario estimar con alta precisión su comportamiento futuro mediante el uso de metodologías robustas que permitieron reducir la incertidumbre del inversionista, minimizar su riesgo de pérdida y gestionar exitosamente su portafolio de inversión.

Al tener riesgos asociados, se requirió gestionar adecuadamente el portafolio de inversión, para poder anticipar variaciones en los mercados financieros que pueden afectar negativamente la rentabilidad, como también influir positivamente en ella. Una forma adecuada de gestionar los efectos de estas variaciones es diversificando las inversiones, haciendo uso de los pronósticos sobre la volatilidad de estos, favoreciendo la reducción del riesgo de pérdida, que como se mencionó previamente, es el interés fundamental de un inversionista.

1.2. FORMULACIÓN DEL PROBLEMA

Partiendo de la premisa de gestionar de manera exitosa un portafolio de inversión en activos financieros, se planteó la siguiente pregunta: ¿Se logran pronósticos de bajo error en la estimación de la volatilidad de series de activos financieros haciendo uso de técnicas de Machine Learning?

Para abordar esta pregunta, debemos tener en cuenta los siguientes aspectos:

- ¿Cuáles son las técnicas de recolección, análisis y exploración de datos aplicadas en las series de tiempo financieras de interés?
- ¿Cómo seleccionar el portafolio de activos con las características adecuadas en torno a su volatilidad y rendimiento?
- ¿Qué técnica de Machine Learning de las empleadas para la estimación de la volatilidad, se ajusta mejor al comportamiento de los activos escogidos?
- ¿Cómo definir la diversificación de una inversión (distribución porcentual de la inversión en los diferentes activos seleccionados) para minimizar el riesgo de pérdida?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Estimar la volatilidad de activos financieros mediante el uso de metodologías de Machine Learning, con el fin de gestionar un portafolio de acciones que reduzca el riesgo de pérdida frente a una inversión.

2.2 OBJETIVOS ESPECÍFICOS

1. Modelar mediante el uso de metodologías de Machine Learning el comportamiento volátil de los activos financieros en los siguientes 90 días.
2. Evaluar la calidad de las predicciones realizadas a partir del cálculo de métricas de ajuste que cuantifiquen la diferencia entre el valor real y el valor predicho de la volatilidad de los activos financieros.
3. Seleccionar la metodología de Machine Learning más eficiente para la estimación de la volatilidad de los activos financieros de interés, con base a las métricas de ajuste establecidas.
4. Determinar la selección de un portafolio de activos financieros factible que garantice una diversificación eficiente, rentable y que minimice el riesgo de pérdida.
5. Desarrollar un dashboard que permite analizar el comportamiento de los activos financieros en los siguientes 90 días y simular la diversificación ideal de una inversión con bajo riesgo de pérdida.

3. MARCO TEORICO Y ANTECEDENTES

3.1. MARCO TEÓRICO

3.1.1. Marco Teórico Contextual

Se considera como propósito fundamental a la hora de crear un portafolio de inversión que se puede llegar a un equilibrio sustancial entre el retorno y el riesgo a la exposición. Los portafolios de inversión a evaluar en este documento están compuestos por activos financieros, los cuales se definen como "un instrumento financiero a los emitidos por las unidades económicas de gasto, que constituyen un medio de mantener riqueza para quienes los poseen y un pasivo para quienes los generan" [1]

En particular los activos seleccionados, son de renta variable, se denominan de esta manera ya que, ya que la rentabilidad de los títulos depende del desempeño de la compañía y las variaciones que puedan presentarse en la Bolsa, en este caso los pagos futuros y la rentabilidad no se visualizan de forma tan tangible como en los activos de renta fija.

Los activos de renta variable, más específicamente las acciones, constituyen los *mercados de capitales* en donde se negocian activos financieros que tienen una duración más allá de un año, casos como por ejemplo los bonos y las acciones.

La escogencia de los activos financieros debe basarse en tres tópicos, la liquidez, la rentabilidad y el riesgo. La liquidez se puede definir como la posibilidad que tiene un activo para a partir de él, lograr realizarse un alto número y volumen de transacciones en los mercados organizados y desarrollados. La rentabilidad como aquello que hace alusión a los beneficios que se obtienen de una empresa o de una inversión.[2] "El riesgo es la contingencia o proximidad de un daño"; y matemáticamente se define como la desviación estándar o la volatilidad.

- Cálculo de la rentabilidad

A la hora de calcular la rentabilidad real en títulos como las acciones, cuyas cotizaciones corresponden a periodos de tiempo pequeños (diario), puede evaluarse según Crespo y Fernández [3], el cálculo logarítmico de las relaciones de las cotizaciones de los activos, como se muestra a continuación:

Donde:

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

Donde:

R_t= Rentabilidad del título
P_t = Cotización del activo financiero al final del período
P_{t-1}= Cotización del activo financiero al inicio del período

- Cálculo del rendimiento esperado:

La definición del rendimiento esperado de una cartera se basa en la combinación de las rentabilidades individuales de cada título, ponderadas por sus respectivos pesos o proporciones dentro de la cartera. Esta fórmula se expresa como la suma de los productos entre las rentabilidades de cada activo y su peso en la cartera.

$$r_p = \sum_{i=1}^{N} X_i r_i$$

Donde:
r_p = Rendimiento esperado del portafolio
X_i= Proporción del valor inicial de la cartera invertida en el título i
r_i= Rendimiento esperado del título i

- Cálculo de la rentabilidad final

Después de poder contar con la rentabilidad real y la rentabilidad esperada, es necesario establecer la rentabilidad final generada, es así como a partir de la volatilidad que hace referencia a la diferencia que podría existir entre el rendimiento real y el esperado por el inversionista.

$$\sigma = \sqrt{\left[\frac{1}{n-1} * \sum_{i=1}^{n} (r_i - r)^2\right]}$$

σ= Desviación estándar o volatilidad del título
r_i= Rendimento esperado del título i
n = Número de sesiones en un año. Para datos diarios, n = 252

La siguiente ecuación se puede definir como un cálculo aproximado del riesgo, que no es más que la covarianza de los rendimientos entre los valores i y j.

$$\sigma_p = \left[\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j \sigma_{ij}\right]$$

σ_p= Riesgo del Portafolio
σ_{ij}= Covarianza de los rendimientos entre el valor i y el valor j
$X_i X_j$= Participaciones o ponderaciones del valor i y el valor j en el portafolio

- Bolsa de Valores

La Bolsa de Valores, en el caso de este estudio corresponde a Colombia, como lo define en su portal web,

es el operador líder de la infraestructura tecnológica del mercado de capitales en el país con presencia en toda la cadena de valor, desde la negociación de los títulos hasta su custodia en el servicio de depósito, es de nuestro interés evaluar las acciones, cuya característica primaria puede definirse como la posibilidad de obtener una ganancia. Generalmente es permitido captar recursos financieros de lo que se considera disponible. Así pues, como ya se comentó anteriormente las acciones son un tipo de inversión de capitalización y su rentabilidad es variable.

Las acciones están determinadas por varios valores de interés como son los comportamientos de los precios de la acción, que están dados por el precio de apertura, el precio más alto alcanzado, el precio más bajo alcanzado y el precio de cierre de la acción. Se conocen como: Open, High, Low y Close que se almacenan en un determinado tiempo de análisis.

Desde un punto técnico, la bolsa de valores también se define como un mercado financiero o de capitales organizado, oficialmente regulado. [4]

La Bolsa de Valores de Colombia (BVC), usa como principal índice bursátil el COLCAP, el cual se compone de los 20 emisores y 25 acciones más líquidas del mercado, ponderando las acciones por capitalización de mercado ajustada sin límite de participación, tal como se explica en el portal web de la BVC. En un índice, la proporción que representa cada acción está determinada por el valor de capitalización bursátil ajustada, que se obtiene al multiplicar el precio más reciente por el número de acciones en circulación y el porcentaje de acciones en manos del público. Esto permite reflejar con mayor precisión la importancia de cada compañía en el índice. Es importante aclarar que la canasta del índice COLCAP estará compuesta por mínimo 20 acciones de 20 emisores diferentes.

Para el año 2022, este índice paso de llamarse COLCAP a MSCI COLCAP. La MSCI es una organización que se enfoca en ofrecer datos y análisis detallados sobre los mercados de inversión a nivel global, con el fin de ayudar a los inversionistas a tomar decisiones informadas para sus carteras. Su objetivo es proporcionar información de alta calidad para facilitar la toma de decisiones de inversión estratégicas.

Actualmente, la responsabilidad de la construcción, el cálculo y el mantenimiento del índice COLCAP ha sido asumida por la empresa MSCI en lugar de la BVC. Esta decisión representa una garantía de calidad para los inversionistas internacionales que buscan invertir en este índice. Además, es importante destacar que el nuevo índice considerará al menos 25 valores y 20 emisores, teniendo en cuenta la liquidez de cada acción en circulación y su cuenta flotante. Cabe mencionar que un emisor puede tener más de un título en el índice, como acciones ordinarias y preferenciales.[5]

3.1.2. Marco Teórico Aplicado

La ciencia de datos es una disciplina interdisciplinaria que utiliza técnicas estadísticas, matemáticas, informáticas y de visualización para extraer conocimientos y comprender patrones a partir de conjuntos de datos. Algunas herramientas como el Machine Learning, término fue acuñado por Arthur Samuel en 1959, quien definió ML como un campo de estudio que brinda capacidad de aprendizaje a las computadoras sin ser programado explícitamente. [6]. Uno de los aprendizajes más conocidos es el aprendizaje supervisado, en el cual, se cuenta con una variable respuesta, que más adelante será predicha.

El uso de ML en el desarrollo de modelos requiere un proceso iterativo, que permita ir refinando los

resultados en función de abordar adecuadamente los objetivos planteados, para el desarrollo de esta estrategia se usó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), que involucra seis etapas fundamentales como lo son:

1. Comprensión del problema o negocio
2. Comprensión de datos
3. Preparación de datos
4. Modelado
5. Evaluación del modelo
6. Implementación del modelo

A partir del desarrollo de la metodología previamente mencionada, se evaluaron algunos de los algoritmos utilizados para llevar a cabo este tipo de problemas (forecasting) en el Machine Learning, por los cuales se pueden contrastar el resultado a partir de la capacidad predictiva entregado por cada uno se encuentran el algoritmo SVM, las redes neuronales recurrentes y los modelos de corta memoria (LSTM).

3.1.2.1. Modelos

- Modelo Base

Un modelo base o más conocido como Baseline, puede definirse como un modelo sencillo que sirve de referencia en un proyecto de Machine Learning. Dentro de las funciones principales que tiene un modelo baseline, es ser un comparativo ante la utilización de técnicas más sofisticadas, por ende, carecen de complejidad y se espera que tengan bajo poder predictivo frente al resto de técnicas utilizadas.

Dentro de los beneficios que puede resultar usar modelos baseline se encuentran:

- Sirven como modelo de referencia versus el resto de los modelos entrenados.
- Un modelo baseline permite mejorar el entendimiento de los datos que se están trabajando.

Generalmente en un problema de regresión, una regla común es crear un modelo baseline que prediga la media o la mediana en los datos de entrenamiento. En cambio, para problemas de clasificación, una regla común, es crear un modelo baseline que prediga la frecuencia de la clase que más ocurre, es decir la moda, en los datos de entrenamiento.

Dentro de lo investigado, se encontró que los procesos GARCH de Bollerslev (1986, 1987) son considerados como modelos base (Baseline) para el análisis de pronóstico de varianza condicional. [7]

- Modelos GARCH

Los modelos generalizados autorregresivos heterocedásticos (GARCH), fue dado a conocer en 1991 por Bollersley. En adicional a los modelos ARCH, la estructura de varianza condicional no depende solo del

cuadrado de los errores retrasados, si no de las varianzas condicionales retrasadas p periodos.

Con el fin de evitar una posible pérdida de precisión en la estimación debido al gran número de coeficientes autorregresivos que pueden estar altamente relacionados entre sí, se ha propuesto una parametrización alternativa y restringida. Esta parametrización se basa en un número limitado de parámetros, lo que ayuda a reducir el problema de la multicolinealidad en la estimación de los coeficientes. [8]

- Modelo GARCH (p, q):

El modelo introducido alrededor del año 90, conocido como $\mathrm{GARCH(p, q)}$ es el siguiente:

$$y_t = \varepsilon_t h_t$$

y_t = Variable dependiente
ε_t = Termino de perturbación
h_t = función de modelación de la varianza condicional

$$h_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \, y_{t-i}^2 + \sum_{j=1}^{p} \beta_j \, h_{t-j}^2 \; donde \; \alpha_i, \beta_j \geq 0 \, , debe \sum_{i=1}^{q} \alpha_i + \sum_{i=1}^{p} \beta_j < 1$$

Donde,

α_0 = Varianza de un periodo de tiempo inicial como término constante.
h_{t-p}^2 = Perturbación al cuadrado de ese periodo de tiempo *(t-p)*
y_{t-q}^2 = Varianza Histórica anterior a ese tiempo *(t-q)*
α_i, β_j = Ponderaciones de la varianza de las observaciones recientes y para la varianza de las observaciones históricas

Las condiciones anteriores permiten que la varianza estimada sea positiva en todos los periodos. En la anterior formulación, se ha supuesto que y_t carece de autocorrelación, así como no se disponen de variables explicativas para la esperanza condicional de este proceso. Entonces la esperanza condicional del modelo $\mathrm{GARCH(p, q)}$ es cero y la varianza:

$$var \, y_t = \frac{\alpha_0}{1 - \alpha(1) - \beta(1)}$$

Cuya distribución es leptocúrtica e incondicionalmente homocedástica, esto indica que la distribución de los datos presenta un comportamiento concentrado alrededor de su valor central y la variabilidad de los datos es constante en toda la distribución, lo cual sería positivo ya que cumple con el supuesto estadístico de que los datos no tienen tendencia ni presentan puntos atípicos.

- Modelos SARIMA

En el caso de que la esperanza de los residuales no fuera 0, es decir que las series evaluadas no cumplieran con el supuesto de ser estacionarias en media, se debe evaluar otra alternativa que permita llegar a la predicción de los retornos por medio del complemento al resultado obtenido por los modelos GARCH(p, q) (encargados de modelar la varianza).

Se conoce la estacionalidad a un patrón que se puede llegar a evidenciar en una serie temporal. Este patrón suele repetirse con una regularidad cada S periodos (periodo de tiempo). Se aborda otro concepto en paralelo que es la diferenciación, cuyo proceso permite eliminar la tendencia de una serie y la diferenciación estacional que permite disminuir en gran medida el comportamiento estacional de la misma. Justo en el caso en el cual se presenta estos dos inconvenientes nace como alternativa la familia de modelos ARIMA estacionales más conocidos como SARIMA. [9]

Definición: Si d y D son enteros no negativos, entonces se dice que $\{Y_t\}$ es un modelo $ARIMA(p,d,q) \times (P,D,Q)_s$ estacional de periodo S si la serie $X_t = (1-B)^d(1-B^s)^D Y_t$ es un proceso $ARMA$ estacionario definido por:

$$\phi(B)\Phi(B^s)X_t = c + \theta(B)\Theta(B^s)\varepsilon_t ,$$

Donde

- $\{\varepsilon_t\} \sim N(0,\sigma^2)$ es ruido blanco,
- $\phi(z) = 1 - \phi_1 - \cdots - \phi_p Z^p$ es el polinomio autorregresivo,
- $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ es el polinomio media móvil,
- $\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P$ es el polinomio autorregresivo estacional,
- $\Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q$ es el polinomio media móvil estacional.
- B es el operador de retraso en el tiempo.
- C representa un término constante o de intercepto que se agrega para modelar el nivel medio o la tendencia de la serie de tiempo.

Definición Componentes Proceso ARIMA:

Componentes no estacionales:

p: Orden del componente AR (autorregresivo), que captura la dependencia lineal de la serie en sus propios valores pasados. Indica el número de términos autoregresivos incluidos en el modelo.
d: Orden del componente I (integrado), que representa el número de diferencias no estacionales necesarias para convertir la serie original en una serie estacionaria.
q: Orden del componente MA (media móvil), que captura la dependencia lineal de la serie en los errores de pronósticos pasados. Indica el número de términos de media móvil incluidos en el modelo.

Componentes estacionales:

P: Orden del componente AR estacional, que captura la dependencia lineal de las series en sus valores pasados en el periodo estacional**.**
D: Orden del componente I estacional, que representa el número de diferencias estacionales necesarias

para convertir la serie en una serie estacionaria en el periodo estacional.
Q: Orden del componente MA estacional que captura la dependencia lineal de la serie en los errores de pronóstico estacionales pasados.

- Support Vector Machine (SVM)

La metodología de Soporte de Máquinas Vectoriales (Support Vector Machine SVM), se puede utilizar para problemas de clasificación y regresión. Es un algoritmo de aprendizaje supervisado y funciona bajo el concepto de cálculo de margen. En este algoritmo, cada elemento de datos se representa como un punto en un espacio n-dimensional (donde n es el número de características que tenemos en nuestro conjunto de datos). El valor de cada característica es el valor de la coordenada correspondiente.

El SVM, clasifica los datos en diferentes clases al encontrar un hiperplano que separa los conjuntos de datos de entrenamiento en clases. Funciona maximizando las distancias entre el punto de datos más cercano (en ambas clases) y el hiperplano, que podemos llamar margen. [10]

El concepto de tratar los objetos para clasificarlos o predecirlos, como puntos en un espacio de alta dimensión y encontrar una línea que los separe no es exclusivo de SVM. El SVM, sin embargo, es diferente de otros clasificadores basados en hiperplanos en virtud de cómo se selecciona el hiperplano.

Si se define la distancia desde el hiperplano de separación al vector de expresión más cercano como el margen del hiperplano, entonces SVM selecciona el hiperplano de separación de margen máximo. La selección de este hiperplano en particular maximiza la capacidad de SVM para predecir la clasificación correcta de ejemplos nunca antes vistos.

El teorema detrás de esta metodología, asume que los datos con los que se entrena la SVM se extraen de la misma distribución que los datos con los que se prueba. Por otro lado, el teorema no asume que los dos conjuntos de datos se extrajeron de una clase particular de distribuciones. Por ejemplo, una SVM no asume que los valores de los datos de entrenamiento siguen una distribución normal. [11]

- Support Vector Regression (SVR)

La regresión de vector de soporte (SVR) tiene como objetivo ajustar un hiperplano a través de un conjunto dado de puntos de datos, donde los puntos pueden estar en el espacio de patrones o en un espacio de características de mayor dimensión.

El objetivo de este ajuste es minimizar la complejidad estructural al mismo tiempo. Determinar tal regresor requiere la minimización de una función cuadrática sujeta a restricciones de desigualdad lineal, que es una tarea de programación convexa. [12]

En comparación con otros regresores de redes neuronales, SVM tiene tres características distintas cuando se utiliza para estimar la función de regresión. Primero, SVM estima la regresión utilizando un conjunto de funciones lineales que se definen en un espacio de características de alta dimensión. En segundo lugar, SVM realiza la estimación de regresión por minimización de riesgos, donde el riesgo se mide utilizando la función de pérdida insensible de Vapnik. En tercer lugar, SVM implementa el principio SRM que minimiza la función de riesgo que consiste en el error empírico y un término regularizado. [13]

Dado un conjunto de puntos (x_1, y_1), (x_2, y_2) … (x_l, y_l) $\left(x_i \in X \subseteq R^n, y_i \in Y \subseteq R\right)$, donde l es el número total de muestras de entrenamiento, aleatorias e independientemente generadas de una

función desconocida, así pues, SVM aproxima la función usando la siguiente forma:

$$f(x) = w \cdot \phi(x) + b$$

Donde $\phi(x)$ representa el espacio finito-dimensional el cual es no linealmente asignado desde el espacio de entrada x. Los coeficientes w y b son estimados por la minimización de la función de riesgo regularizada:

$$\frac{1}{2}\left|\left|w\right|\right|^2 + C\,\frac{1}{l}\sum_{i=1}^{l} L_{\in}\left(y_i\,,\,f(x_i)\right)$$

$$L_{\in}\left(y_i\,,\,f(x_i)\right) = \{|y_i - f(x_i)| - \varepsilon,\ \ |y_i - f(x_i)| \geq \varepsilon\ ,\ 0\ en\ otro\ caso\}$$

El primer termino $\left|\left|w\right|\right|^2$ es llamado del término de regularización. Minimizando $\left|\left|w\right|\right|^2$ hará que la función sea lo mas plana posible. El segundo término $\frac{1}{l}\sum_{i=1}^{l} L_{\in}\left(y_i\,,\,f(x_i)\right)$ es el error empirico medido por la función de perdida $L_{\in}\left(y_i\,,\,f(x_i)\right)$. Esta función de perdida provee ventajas del uso del espacio de los datos puntuales representados en la primera función. C se refiere a la constante de regularización. Este parámetro es definido empiricamente.

- **Extreme Gradient Boosting (Xgboost)**

El gradient boosting es una poderosa técnica de aprendizaje automático basado en arboles de decisión que se utiliza comúnmente en problemas de regresión y clasificación. Al ser una técnica de ensamble el gradient boosting busca mediante el uso de árboles de decisión, construir diferentes modelos consecutivamente y no de forma individual para generar mejores predicciones sobre la variable respuesta. Los nuevos aprendices se construyen con el fin de minimizar la función de perdida asociada a todo el conjunto construido, permitiendo que el algoritmo a partir de los mismos errores generados pueda aprender de ellos.[14]

Existe un conjunto de entrenamiento $\{(x_i, y_i)\}_{i=1}^{n}$, una función de perdida $L\left(y, F(x)\right)$, número de iteraciones M.

Algoritmo[15]**:**

1. Inicializar el modelo con un valor constante:

$$F_0(x) = argmin_{\gamma} \sum_{i=1}^{n} L(y_i\,, \gamma)$$

2. Para $m = 1\ hasta\ M$:
   1. Calcular los llamados pseudo-residuales

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} para\ i = 1, .., n.$$

2. Ajustar un "base learner", por ejemplo, un árbol $h_m(x)$ a los pseudo-residuales. Por ejemplo, entrenándolos usando el conjunto de entrenamiento $\{(x_i, r_{im})\}_{i=1}^{n}$
3. Calcular γ_m resolviendo el siguiente problema de optimización de una dimensión:

$$\gamma_m = \arg\min \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

4. Actualizar el modelo:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

5. Salida $F_M(x)$.

Xgboost utiliza una regularización avanzada que mejora la generalización del modelo y que ayuda en problemas tan frecuentes que suelen presentarse en modelos de Machine Learning como el sobreajuste. Entre otras ventajas que ofrece este tipo de metodologías es importante resaltar la posibilidad del procesamiento paralelo, lo que permite que haya una compensación en la demanda computacional de este algoritmo versus otros.

A la hora de llevar a cabo la ejecución de esta metodología es importante inicializar algunos parámetros denominados hiperparametros y que definirán el modelo:

- Hiperparametros generales: están relacionados con el funcionamiento del algoritmo y define el booster que se usara.
- Hiperparametros booster: Están relacionados al tipo de booster escogido, un ejemplo de estos es la profundidad, el número de árboles, entre otros.
- Hiperparametros de aprendizaje: Son aquellos que permiten definir el escenario de aprendizaje asociado a la función de perdida.

- Redes Neuronales Recurrentes

Las redes neuronales recurrentes, son sistemas dinámicos que hacen un uso eficiente de la información temporal, tanto para la clasificación, como para la predicción [16]. Esto significa que después del entrenamiento, las interrelaciones entre la entrada actual y los estados internos se procesan para producir la salida y representar la información pasada relevante en los estados internos. Durante un proceso de aprendizaje supervisado, los valores objetivo constituyen una segunda fuente de información. Estos valores objetivo especifican las interrelaciones relevantes en la secuencia de entrada. Como respuesta, si es usado para predicción, se obtiene una nueva serie temporal. Dentro de las redes neuronales recurrentes, se pueden hallar metodologías como por ejemplo los modelos LSTM. [16]

- Modelos LSTM (Long Short-Term Memory)

Los modelos de memoria a corto plazo son un tipo de red neuronal recurrente que, aunque entre muchas bondades relacionadas al aprovechamiento de la información que puede ser generada por los datos secuenciales en una serie de tiempo, pueden presentar inconvenientes tales como el tiempo de ejecución

al volverse un algoritmo de aprendizaje lento.

Los modelos LSTM, fueron creados hacia 1990, justamente para resolver este tipo de problemas que podían presentar las redes neuronales recurrentes. Dentro de las etapas o momentos que se deben destacar dentro de esta metodología existen tres.

La primera se considera *"puerta de entrada"* que actúa como un filtro preliminar de la información que puede tenerse en cuenta para ser considerada dentro de la serie y que a su vez determina si hay cabida a una nueva entrada o no. Como paso siguiente, nos encontramos con la *"puerta de olvido"*, que permite identificar qué información es relevante y cual definitivamente solo puede generar ruido y por ende no debe ser tenida en cuenta y en última instancia la *"puerta de salida"*, que es aquella que decide qué información generar, en pocas palabras, se puede definir como un perceptrón de multicapa simple. Estos tres momentos analógicos están basados en una función sigmoidea, con un rango entre 0 y 1. [17]

Definición de funciones:

- Puerta de entrada: $i_t = \sigma(W_i h_{t-1} + W_i h_t)$
- Puerta de olvido: $f_t = \sigma(W_f h_{t-1} + W_f h_t)$
- Puerta de salida: $o_t = \sigma(W_o h_{t-1} + W_o h_t)$
- Estado celda intermedia: $\overline{C} = \tanh(W_c h_{t-1} + W_c h_t)$
- Estado celda (Próxima entrada de memoria): $c_t = (i_t \times \overline{C}) + (f_t \times c_{t-1})$
- Próximo estado: $h_t = o_t \times \tanh(c_t)$
- X_t Vector de entrada
- h_t Vector de salida
- $W, U, y\ f$ Parámetros de matrices y de vector

⊗ Multiplicación de elementos

⊕ Suma/ Concatenación de elementos

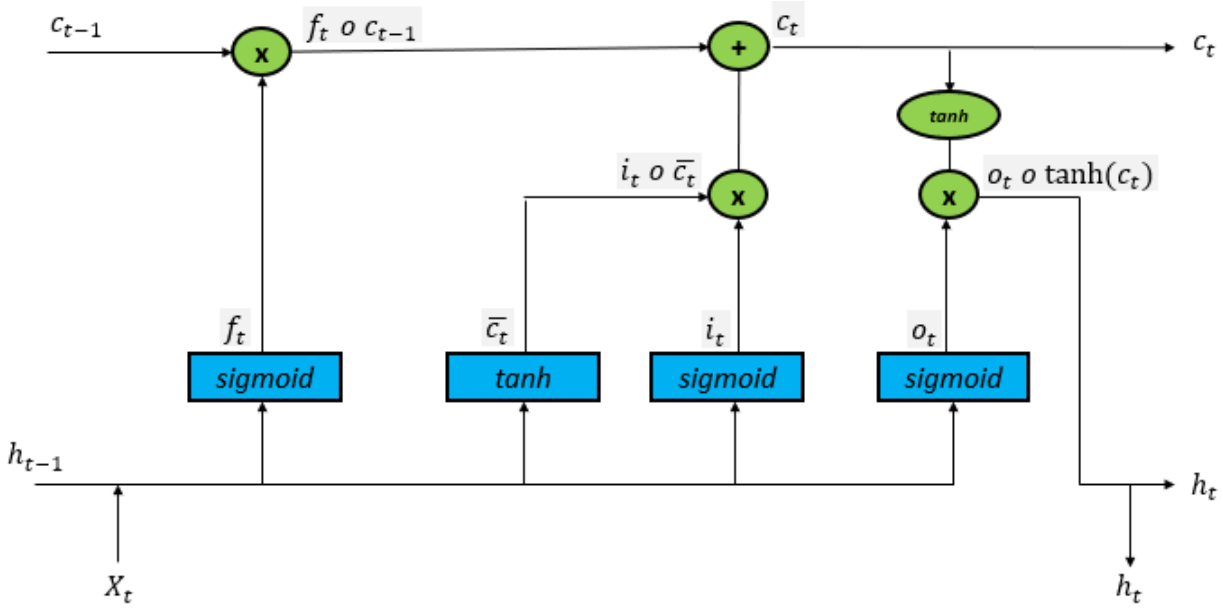


Figura 1 Diagrama de una celda LSTM [18]
Fuente: Elaboración Propia

La ganancia al utilizar este tipo de metodologías es que se puede aprovechar información inclusive ante largos retrasos.

3.1.2.2. Gestión de Portafolios de inversión

La teoría del portafolio se desarrolla en un ambiente de incertidumbre, es decir, los inversionistas no conocen con certeza el verdadero resultado que puede obtenerse al realizar una inversión en un activo o portafolio de activos financieros.

La teoría del portafolio parte del supuesto de que todos los riesgos a los que se expone el inversionista pueden medirse mediante la variabilidad del rendimiento de los activos del portafolio.[19]

Un portafolio puede definirse matemáticamente como una combinación lineal de los activos que contiene. Si se tienen $A_1, A_2, \ldots, A_m$ activos y $w_1, w_{2,\ldots}, w_m$ que indican los pesos o las fracciónes del capital total del portafolio asignadas a cada activo correspondiente. Un portafolio puede definirse como:

$$P = w_1 A_1 + w_2 A_2 + \cdots + w_m A_m$$

Con base a esto, el rendimiento de un portafolio que contiene *m* activos es el promedio ponderado de los rendimientos esperados de los *m* activos, dicho rendimiento está dado por:

$$\overline{r}_P = W' \overline{R}$$

Donde

$\overline{r}_P$= Esperanza matemática del rendimiento de un portafolio.

W'= Vector transpuesto que contiene la participación de cada activo individual en el portafolio.

$\overline{R}=$ Vector de rendimientos esperados individuales

- Selección del portafolio

De acuerdo con el axioma de la media y la varianza, los inversionistas evalúan los activos del portafolio de acuerdo con su rendimiento esperado y su varianza. El inversionista sabe que invirtiendo en un mayor número de activos puede reducir la varianza del portafolio sin disminuir el rendimiento esperado (diversificación útil).

La diversificación eficiente del portafolio o modelo de Markowitz también tiene en cuenta la covarianza o correlación entre los activos del portafolio de tal manera que un portafolio constituido por un gran número de activos correlacionados negativamente puede generar una mayor reducción del riesgo que un portafolio con activos correlacionados positivamente.

El modelo de portafolio de Markowitz no establece sobre cuántos o cuáles activos invertir; el modelo solo establece qué cantidad (proporción de la inversión total) debe invertirse en cada activo, una vez que se hayan definido los activos que conformarán el portafolio.

El modelo de diversificación del portafolio busca determinar la proporción de activos individuales con la combinación riesgo-rendimiento esperado que haga máxima la utilidad esperada del inversionista, definido en el contexto del axioma de la media-varianza. Esto es:

$$\text{Maximizar: } E(U) = f(\overline{r_P}, \sigma_P^2)$$

Donde

$E(U)$= Utilidad esperada del inversionista.

$f(\overline{r_P}, \sigma_P^2)$= Función que relaciona el rendimiento esperado del portafolio $\overline{r_P}$ y la varianza del portafolio σ_P^2.

- Determinación de la frontera eficiente

La frontera eficiente es el conjunto de combinaciones de puntos que representan la mínima varianza posible para un nivel específico de rendimiento esperado. También puede entenderse como el conjunto de puntos en los que se logra un rendimiento óptimo para un nivel determinado de riesgo.

La frontera eficiente se construye a partir de los rendimientos esperados y la varianza de cada uno de los posibles portafolios. El inversionista puede seleccionar, en la frontera eficiente, la combinación de rendimiento esperado y de riesgo según su perfil. La construcción del portafolio eficiente anterior no contempla la no negatividad de las ponderaciones, por lo que podrían obtenerse algunas ponderaciones negativas.

3.1.2.3. Métricas de evaluación de los modelos:

Estas métricas son utilizadas con la finalidad de realizar comparaciones entre distintos modelos ajustados. En el desarrollo del trabajo vamos a evaluar los distintos modelos ajustados haciendo uso de dos métricas

- Raíz del error cuadrático medio (RMSE)
- Error Absoluto medio (MAE)

Wilmot and Matsuura (2005) [20] sugiere que el RMSE no suele ser un buen indicador del desempeño promedio del modelo y podría ser un indicador engañoso del error promedio y, por lo tanto, el MAE sería ser una mejore métrica bajo este propósito. Para la actual propuesta, el indicador del MAE puede ser inclusive una buena opción para la variable a evaluar (Retorno).

Lo anterior no implica centrar la atención en solo esta métrica, si no por el contrario, poder tener la oportunidad de evaluar varias al tiempo y tomar una decisión en conjunto que permita hacer una mejor selección del modelo.

- Raíz del error cuadrático medio RMSE:

Esta métrica realiza una diferencia entre los valores estimados por el modelo y las observaciones reales, estas diferencias se elevan al cuadrado y se calcula un promedio de dichas diferencias. Luego a este indicador se le calcula la raíz cuadrada y permite cuantificar la magnitud del error.

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

Donde y_i es la observación real, $\hat{y}_i$ es el valor estimado y N el tamaño de la muestra.

- Error Absoluto medio (MAE)

Esta métrica representa la diferencia entre los valores estimados por el modelo y las observaciones reales, se calcula esta diferencia como valor absoluto y es dividido sobre el total de los datos, que permite referirnos a esta métrica como un evaluador del promedio de los errores.

3.2. ANTECEDENTES

- **Predicción de precios de cierre de acciones en la bolsa de valores mediante técnicas de aprendizaje automático y minería de datos: evaluación de escenarios para transacciones intradía.** [21]

Guerra, en su trabajo de grado para optar al título de maestría, planteó la posibilidad de hacer la integración de metodologías avanzadas de Machine Learning como Random Forest y Bosques Aleatorios, al análisis de series de temporales para la predicción del precio de cierre de acciones de la bolsa de valores, lo cual nos permite tener un punto de referencia de retar los modelos línea base con metodología de Machine Learning.

Si bien la modelación de este trabajo se realiza en diversas temporalidades, frente a la estimación diaria que se realiza en este trabajo, permite referenciar como realizar el cálculo logarítmico de la diferencia de los precios de cierre.

Entre las conclusiones relevantes de Guerra, ratifica que estas nuevas aplicaciones de Machine Learning pueden llegar a ser tan o más competitivas que las metodologías estadísticas tradicionales.

- Estructuración de un portafolio de inversiones con acciones colombianas [22]

Delgado y Durango, plantearon en su estudio la estructuración de un portafolio de inversión de tres fondos de inversión colectiva del mercado colombiano. Dentro de este ejercicio, hicieron uso de un modelo de optimización que buscaba, cuya función objetivo es minimizar la varianza del portafolio y a su vez la optimización de rendimientos de los tres fondos versus lo reportado por el índice COLCAP. El rendimiento se evaluó a partir del índice Modigliani y Modigliani (M2).

El método de optimización a partir del índice Modigliani y Modigliani (M2) y la teoría de Markowitz son dos enfoques diferentes para construir un portafolio de inversión óptimo. Ambos buscan encontrar una combinación de activos que maximice el retorno esperado dado un nivel de riesgo específico.

La principal diferencia entre los dos métodos radica en cómo manejan la selección del portafolio óptimo. La teoría de Markowitz utiliza la varianza y la covarianza de los rendimientos para calcular la frontera eficiente y encontrar el portafolio óptimo. Por otro lado, el método M2 utiliza la relación entre el retorno y el riesgo (conocido como el coeficiente de Sharpe) para seleccionar el portafolio óptimo.

Una de las principales ventajas del método M2 es que permite la inclusión de activos que no están correlacionados con otros activos en el portafolio. Esto es útil en situaciones en las que la diversificación es difícil de lograr, como en mercados emergentes o en industrias específicas, sin embargo, esto revalida el uso de la Teoría de Markowitz para el contexto de este documento, siendo ampliamente utilizada y reconocida como una de las mejores formas de construir un portafolio óptimo. La teoría de Markowitz considera la diversificación del portafolio y la relación entre el riesgo y el retorno, lo que puede ser especialmente útil para inversores que buscan equilibrar su riesgo y retorno. Además, la teoría de Markowitz puede ser más adecuada para inversores que tienen una visión a largo plazo y pueden tolerar una mayor volatilidad en sus inversiones.

- Optimización financiera de portafolios de inversión en la Bolsa de Valores de Colombia mediante el uso de RGOP [23]

Por su parte Cuartas, buscaba pronosticar la tasa de crecimiento optima (growth optimal portfolio GPO) mediante la aplicación de técnicas de Machine Learning con la intensión de sugerir un portafolio de inversión para acciones en la Bolsa de Valores de Colombia. Como técnica de optimización hizo uso de RGOP (robust growth optimal portfolio) que consiste minimizar el peor de los casos para el valor de riesgo (VaR) de la tasa de crecimiento del portafolio.

Para la selección de las acciones se tuvieron en cuenta aquellas que presentaran los mayores volúmenes de negociación adicional a esto se consideraron aquellas acciones que el mercado podía reconocer, porque generaron los mejores rendimientos y su liquidez era casi segura, se excluyeron acciones que hicieran parte de un mismo grupo empresarial y por último se consideraron acciones que si bien no habían estado en los primeros puestos de la composición del COLCAP, sí habían presentado un crecimiento en la ponderación durante los tres años, dando un total de 8 acciones a considerar.

El criterio de selección mencionado previamente es ajustado en este trabajo teniendo en cuenta criterios de menor varianza obteniendo 10 de las 21 acciones analizadas, de las cuales se realizarán 252 iteraciones de portafolios óptimos compuestos por 5 de las 10 seleccionadas. Los activos a considerar son: Banco de Bogotá, BVC, Celsia, CNEC, Corficolombiana, Grupo AVAL, Nutresa, Preferencial AVAL, Preferencial Corficolombiana y Preferencial Davivienda.

- Construcción de Portafolios de inversión [24]

Uno de los enfoques prácticos populares es la estrategia de reasignación (Wild 2008) adoptada por muchos fondos de pensiones (por ejemplo, Life Strategy Equity Fund, Vanguard). Este enfoque construye una cartera invirtiendo solo en acciones y bonos. Una cartera típica de riesgo moderado, por ejemplo, comprendería un 60% de acciones y un 40% de obligaciones, y la cartera debe reequilibrarse solo semestral o anualmente para mantener esta proporción de asignación. El método ofrece un buen rendimiento a largo plazo; sin embargo, la relación de asignación fija significa que los inversores que prefieren poner más peso en las acciones necesitan tolerar reducciones potencialmente grandes durante los mercados estables.

La estrategia de reasignación mencionada es una forma práctica y simple de construir una cartera de inversión, y se basa en una asignación fija de activos en función del perfil de riesgo del inversor. Sin embargo, este enfoque no tiene en cuenta la diversificación de la cartera, es decir, no considera la correlación entre los diferentes activos, lo que puede llevar a una exposición excesiva a ciertos sectores o a ciertas acciones en particular.

Por otro lado, la teoría de Markowitz tiene en cuenta tanto la rentabilidad esperada como el riesgo de los activos individuales y la correlación entre ellos, lo que permite una construcción de cartera más eficiente y diversificada. Además, la teoría de Markowitz ofrece una frontera eficiente de cartera, que muestra la combinación óptima de activos que maximiza la rentabilidad esperada para un nivel de riesgo determinado.

- Deep Learning y Optimización de Portafolios [25]

En este artículo, los autores adoptan modelos de Deep Learning (aprendizaje profundo) para optimizar directamente el índice de Sharpe de la cartera. La metodología que exponen permite eludir requisitos para pronosticar los rendimientos esperados y les permite optimizar directamente los pesos de la cartera mediante la actualización de los parámetros del modelo. En lugar de seleccionar activos individuales, negocian fondos cotizados en bolsa de índices de mercado para formar una cartera. Los índices de diferentes clases de activos muestran correlaciones sólidas, y negociarlos reduce sustancialmente el

espectro de activos disponibles entre los que elegir. Los autores comparan su método con una amplia gama de algoritmos, y los resultados muestran que el modelo obtiene el mejor rendimiento durante el período de prueba de 2011 hasta finales de abril de 2020, incluidas las inestabilidades financieras del primer trimestre de 2020. Se realiza un análisis de sensibilidad. incluido para aclarar la relevancia de las características de entrada, y los autores estudian más a fondo el rendimiento de su enfoque bajo diferentes tasas de costo y diferentes niveles de riesgo a través de la escala de volatilidad.

Dentro de las metodologías de Deep Learning empleadas, están por ejemplo las redes neuronales recurrentes, más específicamente la metodología de LSTM. El modelo construido consta de tres bloques de construcción principales: capa de entrada, capa neuronal y capa de salida. La idea de este diseño fue utilizar redes neuronales para extraer características transversales de los activos de entrada. Se encontró que las características extraídas de los modelos de aprendizaje profundo rendían mejor que las metodologías tradicionales, siendo esta una de las metodologías a considerar dentro de las evaluadas en este trabajo.

4. METODOLOGÍA

A continuación, se presenta la metodología empleada para realizar la predicción de los rendimientos y la gestión de portafolios de los activos financieros de la BVC. Inicialmente se describe el proceso de consolidación de la información, continuando con la descripción del proceso para el ajuste de los diferentes modelos considerados dentro de proyecto. Adicionalmente se presenta de manera clara y detallada el proceso de validación, selección y comparación de los modelos de Machine Learning vs el modelo base en la proyección de las series financieras.

4.1. Desarrollo de la metodología

En esta sección se especifica en detalle las actividades que conforman la metodología especificando los resultados que se deberían entregar al término de cada una de ellas.

Índice		Tarea	Resultado
1	Consolidación de la información	Recopilación de los datos	• Pre - Master Data Table (Pre-MDT)
		Descripción y exploración de los datos	• Informe de análisis exploratorio de los datos
		Validación de la calidad de los datos	• Reporte sobre la calidad de la información
2	Preparación de la base de datos	Selección de los datos	• Motivos de exclusión • Periodo de tiempo considerado
		Preprocesamiento	• Resultados del preprocesamiento
		Construcción de variables adicionales	• Atributos generados • Máster Data Table (MDT)
3	Proceso de Modelación	Construcción del modelo	• Modelo Base • Modelos ML • Configuración de hiperparametros • Predicciones
		Evaluación de los modelos	• Estrategia de Validación • Métrica • Resultado del modelo
4	Evaluación de los modelos	evaluación de los resultados	• Mejor Modelo según cada serie
5	Selección activos potenciales para conformación de portafolios	Evaluación de 252 portafolios factibles, de los cuales se evalúa su distribución para minimizar su riesgo	• Pesos de distribución de la inversión entre 5 acciones que conforman cada uno de los 252 portafolios factibles.
6	Gestión de portafolio	Definición del portafolio de inversión	

Tabla 1 Relación Objetivo, Tarea Desarrollada y Resultado

4.2. Consolidación de la información

El proceso de descarga de la información de interés corresponde a los activos de renta variable que conforman el índice COLCAP, las cuales se listan a continuación:

COMPAÑIA	SIMBOLO	INDUSTRIA
Bancolombia	BCOLOMBIA	Banca
Banco de Bogotá	BOGOTA	Banca
Bolsa de Valores de Colombia	BVC	Bolsa de valores
Celsia	CELSIA	Energía
Cementos Argos	CEMARGOS	Materiales de construcción
Cemex Latam	CLH	Materiales de construcción
Canacol Energy	CNEC	Petróleo
Corficolombiana	CORFICOLCF	Servicios financieros
Ecopetrol	ECOPETROL	Petróleo
Almacenes Éxito	EXITO	Retail
Grupo Argos	GRUPOARGOS	Conglomerado
Grupo Aval	GRUPOAVAL	Banca
Grupo Sura	GRUPOSURA	Conglomerado
Interconexión Eléctrica	ISA	Energía
Grupo Nutresa	NUTRESA	Industria alimenticia
Grupo Aval	PFAVAL	Banca
Cementos Argos	PFCEMARGOS	Materiales de construcción
Corficolombiana	PFCORFICOL	Servicios financieros
Davivienda	PFDAVVNDA	Banca
Grupo Argos	PFGRUPOARG	Conglomerado
Grupo Sura	PFGRUPSURA	Conglomerado
Pacific Exploration & Production	PREC	Petróleo

Tabla 2 Descripción de series de las acciones consideradas
Elaboración propia

La información se toma del portal de la Bolsa de Valores de Colombia [1], en el periodo comprendido entre

01/10/2019 a 30/09/2022, para un total de 36 meses, que contienen 734 días de cotización, donde la volatilidad a estimar será calculada respecto a las variaciones diarias del precio de cierre del activo.

El proceso de extracción de la información conlleva a realizar un barrido por cada activo de interés, obteniendo cortes de seis meses, para un total de 6 extracciones por cada uno de los 22 activos para un total de 132 archivos en formato .csv, de igual manera se consideran 6 archivos adicionales con similares características para el índice COLCAP, que será considerado como variable exógena para robustecer la estimación.

Posteriormente se hace un proceso de consolidación por cada activo de renta variable, donde se ajusta formatos de fecha y se lleva a un esquema con las siguientes características:

Nemotécnico	fecha	Cantidad	Volumen	Precio Cierre	Precio Mayor	Precio Medio	Precio Menor
EXITO	10/21/2019	4,807,634	85,902,992,320	17,800.00	17,880.00	17,868.04	17,800.00
EXITO	10/22/2019	4,045,959	72,250,168,480	17,860.00	17,860.00	17,857.37	17,800.00
EXITO	10/23/2019	1,495,931	26,751,517,140	17,880.00	17,900.00	17,882.85	17,860.00
EXITO	10/24/2019	640,883	11,463,999,780	17,900.00	17,900.00	17,887.82	17,880.00
EXITO	10/25/2019	10,476,170	187,197,588,080	17,860.00	17,900.00	17,868.90	17,860.00
EXITO	10/28/2019	392,745	7,020,044,840	17,880.00	17,880.00	17,874.31	17,860.00
EXITO	10/29/2019	1,021,698	18,247,153,700	17,860.00	17,860.00	17,859.64	17,820.00
EXITO	10/30/2019	406,919	7,273,599,500	17,880.00	17,880.00	17,874.81	17,860.00

Tabla 3 Tabla: Variación diaria precios de cierre Activo EXITO.
Elaboración propia

Los archivos consolidados encuentran almacenados en una carpeta, de donde son cargados al ambiente Python mediante la librería OS[2], para listar los archivos y generar el proceso de lectura de una manera sistemática.

4.3. Preparación de los datos.

En esta parte se construyó la MDT (Master DataTable) final con la que se desarrollaron los modelos y sus respectivas validaciones, ejecutando los siguientes pasos:

1. Se excluyeron del análisis series que no presentaran un retorno diferente a cero en el periodo considerado.
2. Se construyó la variable de respuesta en función de la rentabilidad de acción de la siguiente forma:

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

   Donde P hace referencia al precio de cierre.

3. El periodo considerado para el análisis fue desde el 10 de enero 2019 al 30 de noviembre del 2022.

4.4. Proceso de modelación.

Con la finalidad de realizar la comparación entre modelos tradicionales con los modelos de Machine Learning, se seleccionó un modelo base de series de tiempo y dos modelos de ML.
El modelo de series de tiempo servirá como un punto de partida para la comparación de los modelos restantes, frente a los modelos de Machine Learning se seleccionaron modelos de aprendizaje supervisado enfocados en regresión. En este trabajo se abordaron los siguientes modelos.

- Modelo base: Modelo GARCH
- Support Vector Machine en su versión con enfoque de regresión
- XGBoost
- Redes neuronales recurrentes LTSM

Para el proceso de validación se definieron dos métricas las cuales son el RMSE y el MAE los cuales permiten definir la presión de los modelos, en cuanto más cercanos sean a cero mejor se podría considerar las estimaciones realizadas por este.

Frente al ajuste de los modelos se consideró el siguiente proceso:

1. Se listaron todos los parámetros de los modelos y se definieron un rango de valores a ser considerados para ajustar el modelo.
    a. Modelo GARCH:
        i. Parámetros del modelo de la media mediante ARMA:
            1. p: [1, 2]
            2. q: [1, 2]
        ii. Modelo para la varianza: ['sGARCH', 'eGARCH', 'gjrGARCH', 'iGARCH', 'apARCH']
        iii. Parámetros del modelo de la varianza:
            1. p: [1, 2]
            2. q: [1, 2]
        iv. Distribución: ['norm', 'std', 'ged', 'sstd']
        v. Incluir la media: [True, False]
    b. Support Vector Machine:
        i. Numero de rezagos: [10, 12, 15]
        ii. kernel: [linear, poly, rbf , sigmoid]
        iii. gamma: [0.5, 1, 1.5, 2]
        iv. épsilon: [0, 0.01, 0.05]
        v. Costo: [25, 50, 100, 150]
    c. Redes neuronales recurrentes LSTM:
        i. 'epochs':[100,200,500]
        ii. 'batch_size':15,
        iii. 'validation_split':0.1,
        iv. 'shuffle': False,

v. 'verbose':0,
vi. 'seed':1234
vii. 'learning_rate': [0.01,0.1]
viii. 'drop_out': [0.1, 0.2]

d. XGboost
   i. Técnica Tuneo Hiperparámetros: Grid Search
   ii. Timesteps (‘numero de rezagos’): [5,10]
   iii. N estimators: [100, 500, 900, 1100, 1500]
   iv. Max depth: [2, 5, 15]
   v. Learning rate: [0.05, 0.1, 0.20]
   vi. Min child weight: [1, 2, 3]

Los anteriores rangos de posibles parámetros en la construcción de cada metodología se construyeron teniendo en cuenta a partir de la teoría los valores mínimos y máximos posibles. Además, se buscó que el valor que pudiese ser seleccionado durante la búsqueda de menor error, no afectaran problemas comunes como es el sobreajuste en el Machine Learning.

2. Se definieron un porcentaje de train y test del 90% y 10% respectivamente [26], en especial considerando que se contrastaran métodos de Machine Learning, donde la evaluación del modelo en la base de testeo permitió tener mayor solidez en las comparaciones de las métricas de ajuste que se tendrán en cuenta para la medición de desempeño.

   Para el ejercicio propuesto se tienen series de 36 meses de cotización de activos financieros, con un total de 733 días, por lo que bajo la recomendación de la literatura correspondería a 73 días de validación, sin embargo, con el fin de proveer un tramo de validación más estándar y mayor certeza a los intereses de un inversor instrumentos de renta variable quienes ya tienen algo de conocimiento en la bolsa de valores y han analizado bien sus opciones antes de invertir, por lo que este tiempo es mínimo para mantener su inversión y lograr recuperar los costos derivados de comisiones y gastos de broker o intermediación.

   Para realizar el proceso de validación, se realizarán metodologías de backtesting,[27] el cual consiste en evaluar el comportamiento de un modelo predictivo al aplicarlo de forma retrospectiva sobre datos históricos. Por lo tanto, es una estrategia de validación que permite cuantificar la capacidad predictiva de un modelo, a continuación, se presentan las alternativas consideradas:[28]

- Selección del modelo con mejor ajuste

1. Se construye para cada combinación de parámetros el modelo correspondiente
2. Se calcula las métricas de evaluación definidas, en la muestra de entrenamiento
3. Se seleccionó como modelo inicial para cada metodología aquella que genere el menor RMSE

4. Se repite el proceso para todas las series consideradas.
5. Una vez seleccionado el mejor modelo de cada metodología, en función de las métricas de desempeño en la muestra de validación se escogerá el mejor de las tres metodologías estimadas.

4.5. Definición del portafolio de inversión

Se ha utilizado el lenguaje de programación Python para llevar a cabo un proceso de selección y combinación de activos, con el fin de generar una lista de 252 portafolios factibles. Se seleccionaron los 10 activos de menor varianza y se agruparon en grupos de 5 para estimar la distribución de cada combinación, considerando el perfil de inversionista y buscando una diversificación eficaz. Esto fue necesario debido al reto de contar con 21 acciones, lo que hubiera resultado en la generación de 20,349 portafolios factibles, lo que excede las capacidades computacionales disponibles.

En este caso, se busca construir la frontera eficiente utilizando la teoría de Markowitz para definir el portafolio de inversión. La meta es maximizar el indicador de Sharpe, que pondera la rentabilidad ajustada al riesgo de un portafolio de inversión. Esta medida se utiliza para evaluar la rentabilidad de un portafolio en relación con el riesgo asumido para obtener esa rentabilidad y, de esta manera, determinar cuál es el mejor portafolio de inversión.

La fórmula del índice de Sharpe es la siguiente:

$$Sharpe\ Ratio = \frac{(Retorno\ Portafolio\ -\ Tasa\ Libre\ de\ Riesgo)}{Desviación\ estándar\ del\ portafolio}$$

El índice de Sharpe es una medida que compara la rentabilidad de un portafolio de inversión con la tasa libre de riesgo, que por lo general se estima a partir del rendimiento de los bonos del Tesoro. Esta medida también tiene en cuenta el riesgo que se asume para obtener esa rentabilidad, el cual se mide a través de la desviación estándar del portafolio.

El índice de Sharpe se interpreta como una medida de la rentabilidad ajustada al riesgo del portafolio. Un índice de Sharpe mayor indica un mejor rendimiento ajustado al riesgo. Es decir, que el portafolio ha logrado una rentabilidad superior en relación con el riesgo asumido. Por otro lado, un índice de Sharpe bajo indica que el rendimiento del portafolio no es suficiente para compensar el riesgo asumido. En resumen, el índice de Sharpe proporciona una forma útil de evaluar la eficacia de la gestión de riesgos de un portafolio de inversión.

Así mismo, es importante destacar que el criterio de Sharpe se convierte en una herramienta valiosa para la comparación de diversos portafolios de inversión y para la evaluación de la efectividad de una estrategi a de inversión en términos de riesgo.

Cabe mencionar que la rentabilidad máxima de un portafolio se determina por el producto entre los peso s asignados a cada acción en cada portafolio construido y el valor esperado del rendimiento de cada acció n.

4.6. Construcción de dashboard

Se llevó a cabo la construcción del dashboard utilizando Power BI, lo que permitirá al usuario comparar diferentes activos en términos de su cotización, medidas de volatilidad y horizonte de estimación. Para estimar la volatilidad de cada activo se utilizará el modelo que presente el mejor desempeño siguiendo las pautas previamente descritas. Se generará una predicción para los próximos 90 días y con los resultados obtenidos, el usuario podrá seleccionar el conjunto de activos que desee y obtener la distribución eficiente de la inversión que minimice el riesgo de pérdida.

A continuación, se ilustra la propuesta grafica del dashboard:

- Análisis descriptivo

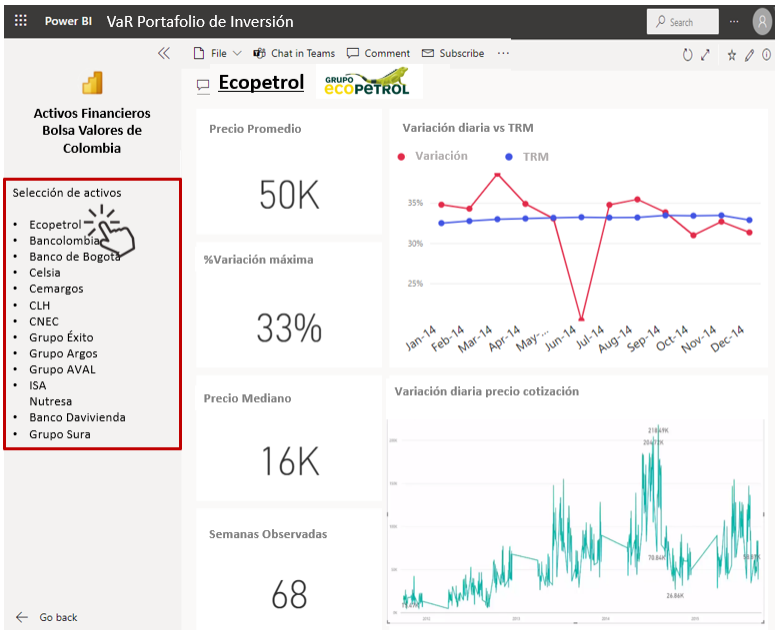


Figura 2 Mockup Tablero Sección: Análisis Descriptivo

- Selección de portafolio eficiente

Figura 3 Mockup Tablero Sección: Selección Portafolio

Tanto la figura 2 y figura 3 representan un mockup (bosquejo inicial) de lo que se pretendía plasmar en el tablero. En la figura 2 se buscó poder entregar al lector la posibilidad de tener el comportamiento descriptivo de las acciones evaluadas. Allí la idea fue plasmar la necesidad de contar con información relevante del precio de cierre, comportamiento a lo largo del tiempo de estos precios, entre otros.

En la figura 3, el objetivo fue plasmar la necesidad de poder seleccionar la cantidad de activos por parte del usuario, para la construcción de su portafolio de inversión, poder visualizar cómo se comporta la frontera eficiente para ese determinado activo y poder evaluar cómo es la distribución porcentual de la inversión en cada activo que compone el portafolio construido.

A continuación, se ilustra el proceso metodológico para facilitar el entendimiento del lector:

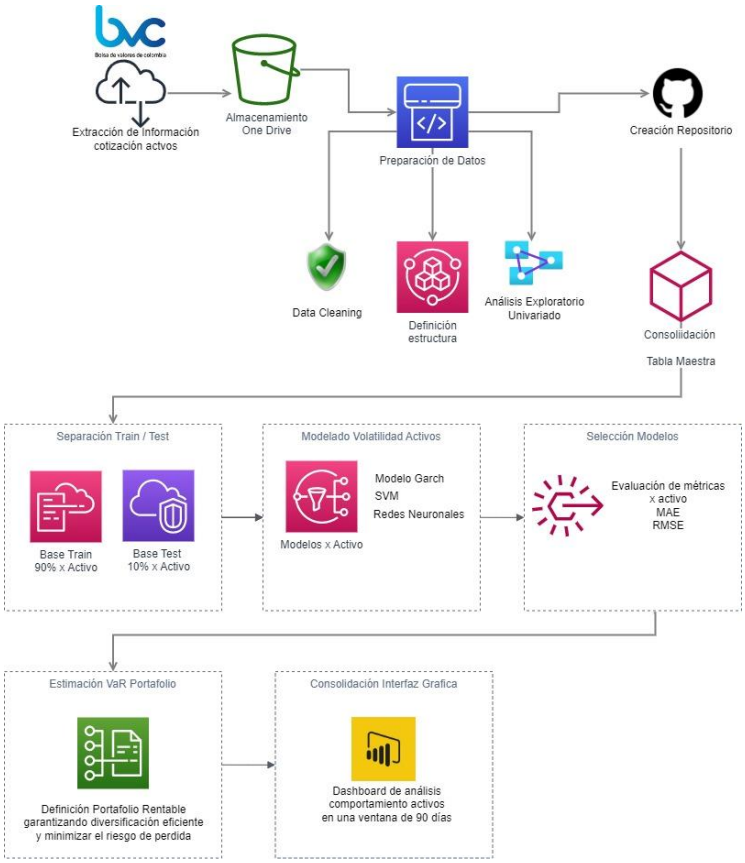


Figura 4 Proceso Metodológico, Fuente: Elaboración propia

5. RESULTADOS

La información utilizada para este proyecto se obtuvo de la bolsa de valores de Colombia, específicamente de las series correspondientes a los emisores participantes en MSCI COLCAP (21 acciones), que incluyen empresas como Bancolombia, Ecopetrol y Grupo Sura, entre otros.

Aunque se listaron en un principio 22 acciones, la acción correspondiente a Pacific Exploration & Production (PREC), fue excluida ya que la variacion en la serie para el tiempo evaluado fue cero.

Se estableció como criterio mínimo para el análisis un período de 4 años para cada acción. Se utilizó una periodicidad diaria para las series y se tomó como variable de interés el precio de cierre de la acción para poder calcular la volatilidad a partir de los precios de cierre de los días t y $t-1$.

5.1 Definición del conjunto de entrenamiento y validación

Se dispuso de un conjunto de datos con 15.393 registros que incluía 21 series de acciones seleccionadas, cada una con 733 observaciones de retornos obtenidos desde el 2 de febrero de 2019 hasta el 30 de septiembre de 2022 provenientes de la bolsa de valores de Colombia.

Para la creación del conjunto de datos de entrenamiento se excluyeron los últimos tres meses,

correspondientes al período del 1 de julio al 30 de septiembre de 2022, que se reservaron para la validación de los diferentes modelos.

Figura 5 Separación de la data set en entrenamiento y test para cada serie de las acciones de la BVC

Descriptivas Iniciales

A continuación, se relaciona descriptivas de tendencia central como el mínimo, el máximo, el promedio y de variación como la desviación estándar tanto de la variable precio cierre como de la variable logaritmo de la variación del precio cierre.

Acción	Min precio cierre	Max precio cierre	Prom precio cierre	Std precio cierre
BCOLOMBIA	17700	45500	32.960	6.433
BOGOTA	28030	89960	69.397	12.935
BVC	7450	12740	10.480	1.219
CELSIA	2813	4870	4.276	349
CEMARGOS	2420	7630	5.328	1.085
CLH	1360	4839	3.239	850
CNEC	6320	13000	10.137	1.058
CORFICOLCF	16500	33600	27.947	3.685
ECOPETROL	1380	3610	2.564	474
EXITO	10000	17980	12.733	1.986
GRUPOARGOS	7010	18980	12.768	2.666
GRUPOAVAL	640	1495	1.061	197
GRUPOSURA	15000	57780	27.047	7.738
ISA	11800	29250	21.232	2.438
NUTRESA	17000	50900	27.457	8.188
PFAVAL	580	1505	1.042	212
PFCEMARGOS	2220	6200	3.999	774
PFCORFICOL	13500	28200	22.850	3.068
PFDAVVNDA	22960	46900	32.636	5.268
PFGRUPOARG	5650	14000	9.431	1.844
PFGRUPSURA	*11200*	*30980*	*20.271*	*4.252*

Tabla 4 Descriptivas Precio Cierre

Acción	Mín log variación precio cierre	Máx log variación precio cierre	Prom log variación precio cierre	Std log variación precio cierre
BCOLOMBIA	-0.196406	0.217478	-0.000343	0.027971
BOGOTA	-0.290619	0.107295	-0.001406	0.023244
BVC	-0.102928	0.095310	-0.000545	0.018068
CELSIA	-0.152526	0.103606	-0.000604	0.017811
CEMARGOS	-0.274437	0.200671	-0.001037	0.028906
CLH	-0.216671	0.185403	-0.000504	0.030985
CNEC	-0.136711	0.147636	-0.000847	0.020267
CORFICOLCF	-0.223144	0.114603	-0.000705	0.019139
ECOPETROL	-0.223987	0.122602	-0.000476	0.028118
EXITO	-0.198451	0.188291	-0.000746	0.025475
GRUPOARGOS	-0.224489	0.287204	-0.000766	0.033447
GRUPOAVAL	-0.165514	0.095026	-0.000911	0.020290
GRUPOSURA	-0.236077	0.290065	0.000174	0.033122
ISA	-0.275626	0.138683	-0.000156	0.024967
NUTRESA	-0.119600	0.218689	0.000520	0.023030
PFAVAL	-0.168501	0.205852	-0.001112	0.021607
PFCEMARGOS	-0.198851	0.162519	-0.001365	0.025241
PFCORFICOL	-0.070404	0.060524	-0.000779	0.016327
PFDAVVNDA	-0.156856	0.124524	-0.000476	0.021205
PFGRUPOARG	-0.265281	0.172633	-0.001102	0.029418
PFGRUPSURA	-0.258235	0.194886	-0.001156	0.027064

Tabla 5 Descriptivas logaritmo de la variacion del precio cierre

Se relaciona gráficamente el comportamiento de algunas de estas series, con el fin de visualizar el comportamiento volátil que se presenta al trabajar con el precio de cierre y el retorno.

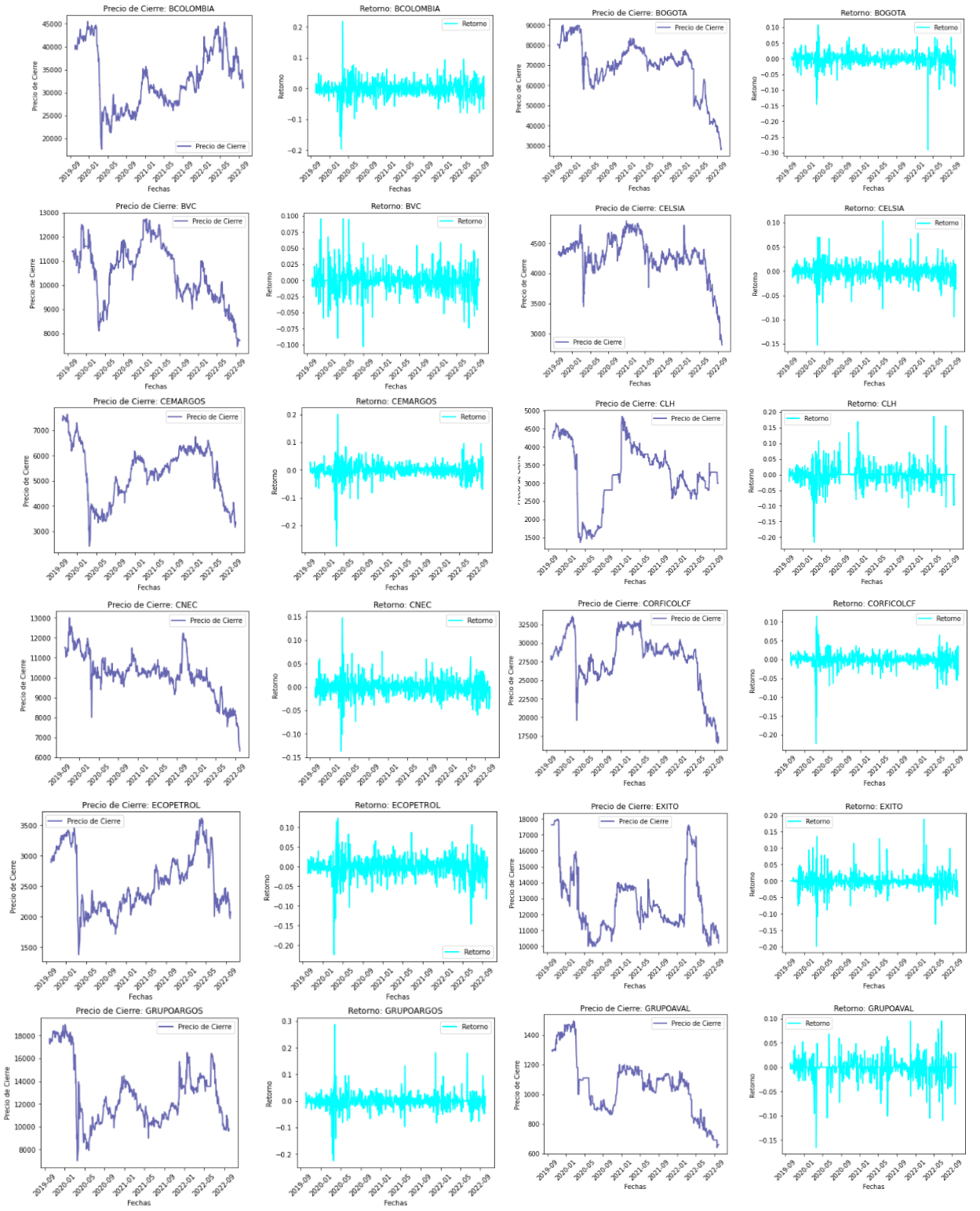


Figura 6 Gráfico Descriptivo: Precio vs Rendimiento (Primeras Acciones)

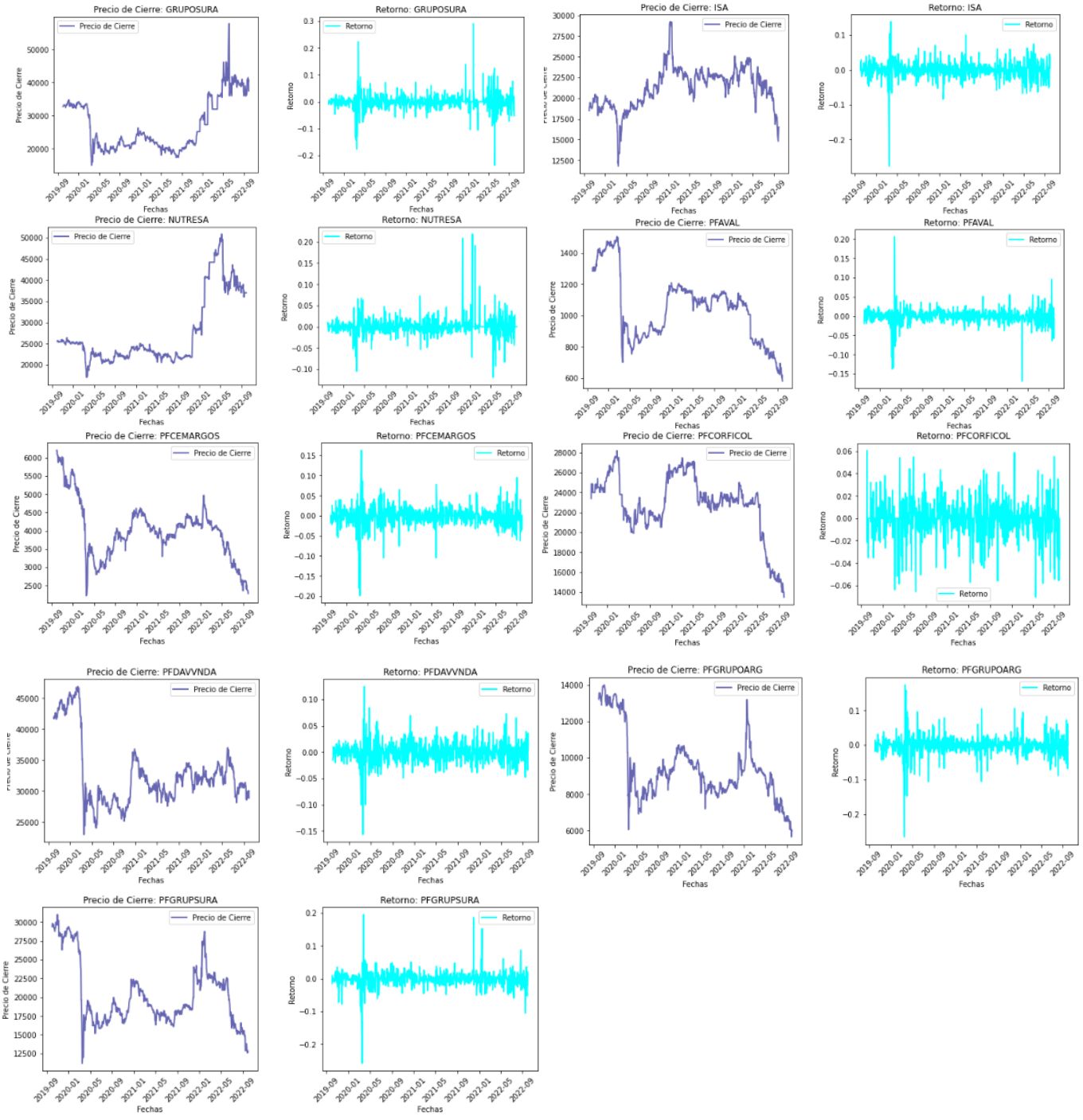


Figura 7 Gráfico Descriptivo: Precio vs Rendimiento (Complemento)

5.2 Estimación de los distintos modelos de Machine Learning

En esta se sección se expondrán los resultados obtenidos de la aplicación de las distintas metodologías escogidas para realizar la proyección de las 21 series.

5.2.1 Estimación del modelo base

La metodología de Box-Jenkins es una técnica de modelización de series de tiempo que involucra tres fases: identificación, estimación y diagnóstico del modelo. En la fase de identificación se busca seleccionar el modelo ARMA y GARCH que mejor ajuste a la serie de tiempo, en la fase de estimación se estiman los parámetros del modelo seleccionado y en la fase de diagnóstico se evalúa si el modelo seleccionado es adecuado para la serie de tiempo en cuestión.

En este caso, se aplicó la metodología de Box-Jenkins para ajustar un modelo ARMA para la media de la serie y un modelo GARCH para la varianza de la serie. Antes de la estimación del modelo, se validó la estacionariedad de la serie mediante la prueba de Dickey-Fuller, encontrando que todas las series consideradas cumplen con este supuesto bajo un nivel de significancia del 5%. Esto implica que la serie de tiempo no presenta tendencia ni cambios en su varianza a lo largo del tiempo, lo que permite aplicar los modelos ARMA y GARCH de manera adecuada.

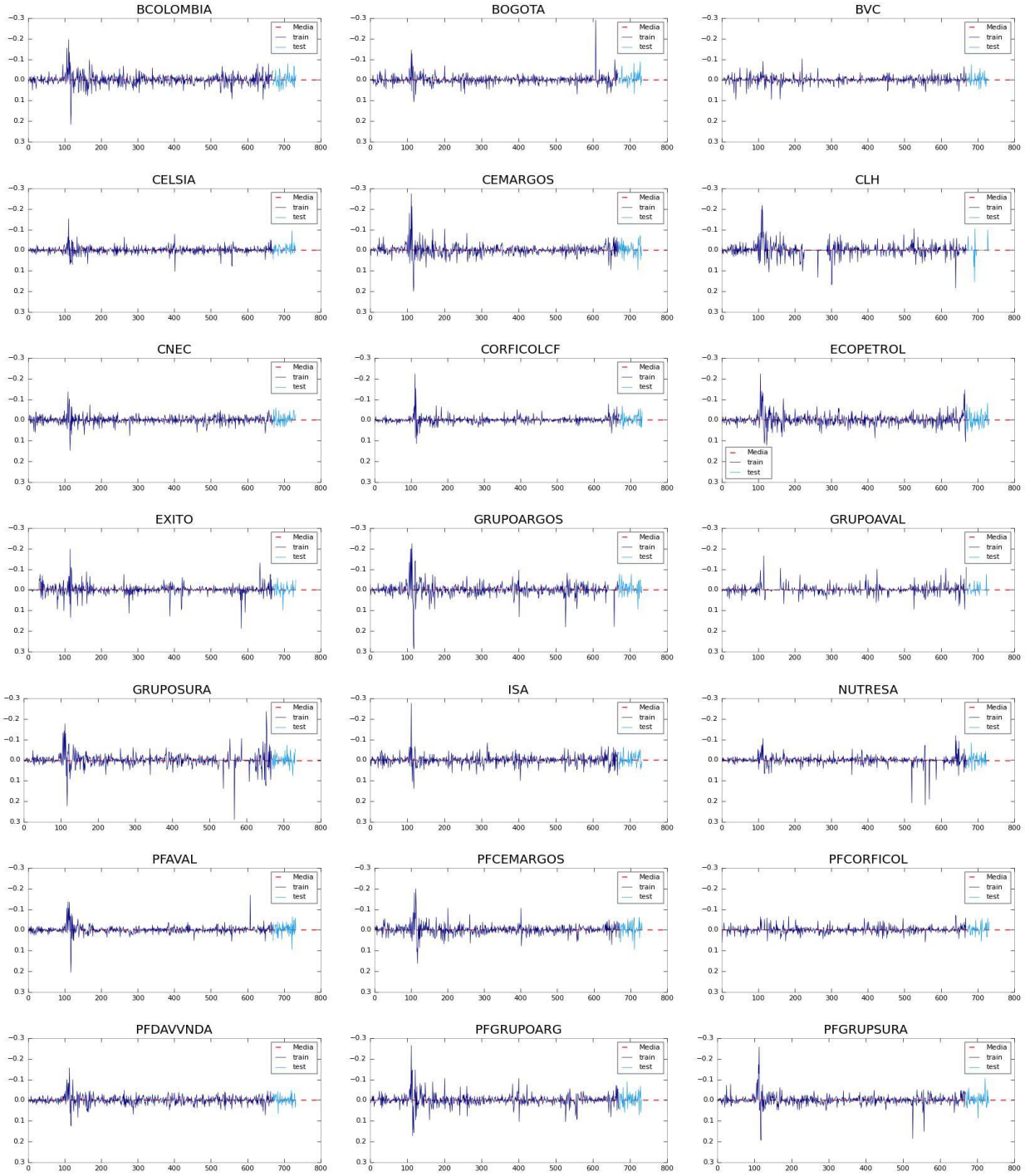


Figura 8 División de la data set en entrenamiento y prueba para cada conjunto de datos

Se procede a hacer la estimación de la media mediante un modelo ARMA y de la varianza mediante un modelo GARCH donde los hiperparametros obtenidos para cada serie fueron los siguientes:

Índice	Acción	ARMA			GARCH				AIC	RMSE	MAE
		p	q	Incluir Media	p	q	Modelo Varianza	Distribución Retornos			
1	BCOLOMBIA	2	2	FALSE	1	1	apARCH	std	-4.759	0.042	0.035
2	BOGOTA	1	1	FALSE	2	1	sGARCH	ged	-6.039	1.991	1.990
3	BVC	1	1	FALSE	1	1	eGARCH	ged	-7.875	5.493	5.015
4	CELSIA	1	1	FALSE	1	1	sGARCH	ged	-5.789	0.029	0.023
5	CEMARGOS	2	2	FALSE	2	1	iGARCH	sstd	-4.990	0.058	0.052
6	CLH	1	1	FALSE	1	1	eGARCH	ged	-9.286	0.048	0.041
7	CNEC	1	2	TRUE	1	2	sGARCH	std	-5.396	0.033	0.028
8	CORFICOLCF	1	1	FALSE	1	1	iGARCH	ged	-6.022	0.042	0.036
9	ECOPETROL	2	2	FALSE	2	2	eGARCH	ged	-4.859	0.049	0.039
10	EXITO	2	1	FALSE	2	1	eGARCH	ged	-9.169	1.008	1.000
11	GRUPOARGOS	1	1	FALSE	1	1	apARCH	sstd	-7.340	9.177	5.723
12	GRUPOAVAL	2	1	FALSE	2	2	apARCH	ged	-15.402	0.029	0.024
13	GRUPOSURA	1	1	FALSE	1	1	apARCH	sstd	-15.944	7.682	4.346
14	ISA	2	2	TRUE	1	1	gjrGARCH	std	-5.040	0.053	0.047
15	NUTRESA	1	1	FALSE	2	2	apARCH	sstd	-11.045	5.601	3.490
16	PFAVAL	1	2	FALSE	2	2	iGARCH	std	-5.764	0.055	0.048
17	PFCEMARGOS	1	1	FALSE	1	2	sGARCH	ged	-6.180	2.224	2.193
18	PFCORFICOL	2	2	FALSE	2	2	eGARCH	ged	-9.468	0.619	0.614
19	PFDAVVNDA	2	2	FALSE	1	1	gjrGARCH	ged	-5.271	0.029	0.024
20	PFGRUPOARG	1	1	FALSE	1	1	gjrGARCH	ged	-5.028	0.088	0.081
21	PFGRUPSURA	2	2	TRUE	1	2	apARCH	sstd	-5.214	0.044	0.037

Tabla 6 Hiperparametros y métricas del modelo base

Es relevante mencionar que los hiperparametros mencionados anteriormente se obtuvieron mediante simulación. Para lograr un mejor ajuste de las series, se amplió el rango de los posibles valores a probar, lo que dio lugar a un total de 640 modelos ajustados a partir de la combinación de dichos parámetros. La métrica utilizada para la selección de los mejores parámetros fue la que presentara el menor valor de AIC (Criterio de Información de Akaike).

En la Figura 8 se puede apreciar el comportamiento del modelo sobre la base de datos de validación. Se puede observar que la estimación realizada mediante este modelo no es buena, ya que le resulta difícil ajustarse a la variabilidad de los retornos de las series. En casos como los de Bogotá y la BVC, los valores pronosticados no se parecen en absoluto a la serie real, lo que se refleja en valores de MAE y RMSE mayores en comparación con otras series estimadas.

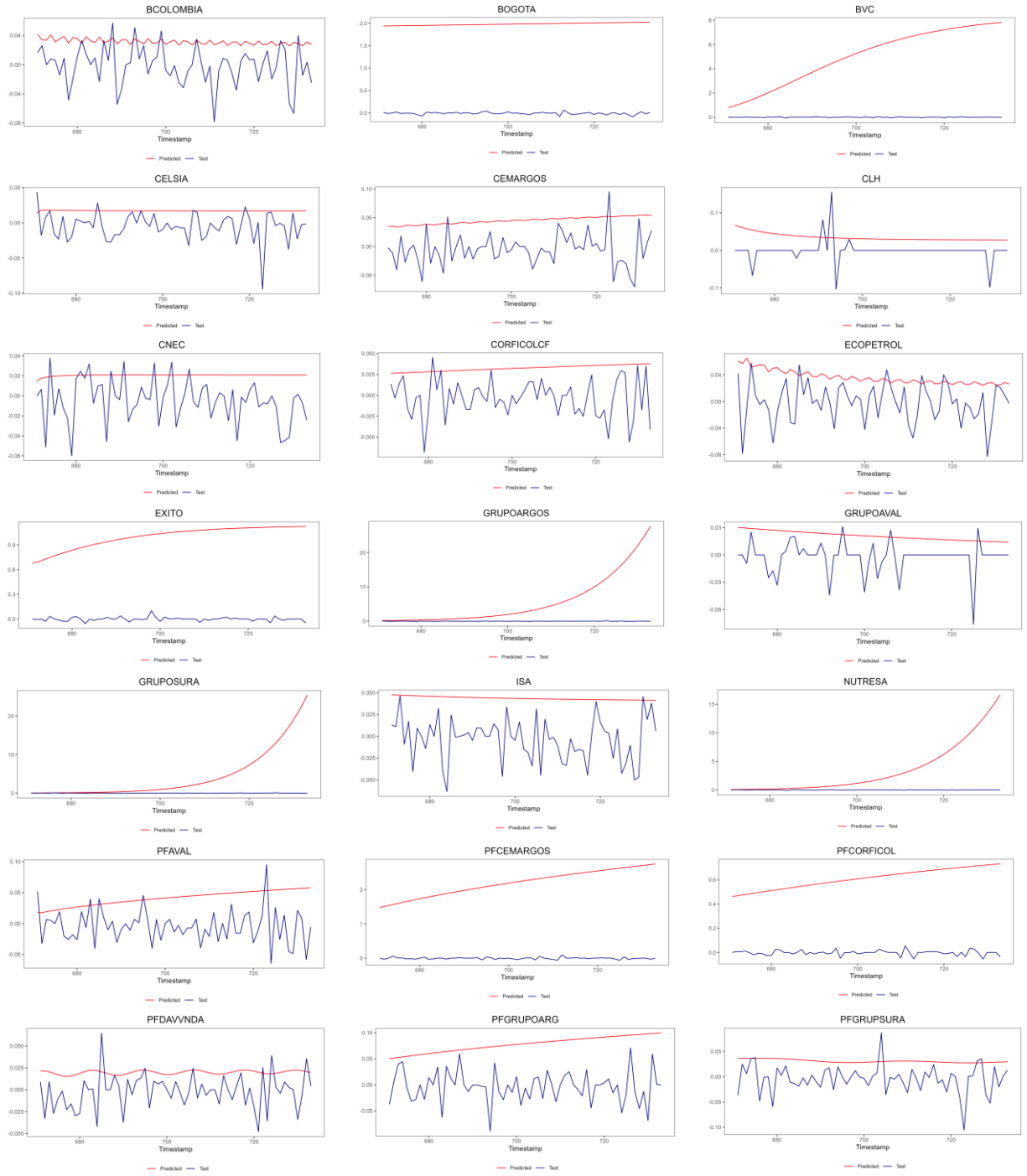


Figura 9 Comparativo serie de prueba Vs Predicción del modelo GARCH

5.2.2 Ajuste del modelo mediante Support Vector Regression

Dado que se consideraron diferentes hiperparámetros para ajustar las series de tiempo utilizando esta metodología, se ajustaron un total de 576 modelos para cada serie y se seleccionó el modelo que presentó un mejor ajuste utilizando el MAE como métrica de selección. Como resultado, se obtuvieron diferentes conjuntos de hiperparámetros para cada serie.

índice	Acción	Núm. Rezagos	kernel	gamma	épsilon	C	RMSE	MAE
1	BCOLOMBIA	10	rbf	0.5	0	25	0.016	0.024
2	BOGOTA	10	rbf	0.5	0	25	0.012	0.021
3	BVC	10	rbf	0.5	0	25	0.010	0.017
4	CELSIA	10	rbf	0.5	0	25	0.009	0.015
5	CEMARGOS	10	rbf	0.5	0	25	0.014	0.023
6	CLH	10	rbf	0.5	0	25	0.014	0.026
7	CNEC	10	rbf	0.5	0	25	0.011	0.016
8	CORFICOLCF	10	rbf	0.5	0	25	0.009	0.015
9	ECOPETROL	10	rbf	0.5	0	25	0.014	0.023
10	EXITO	10	rbf	0.5	0	25	0.012	0.024
11	GRUPOARGOS	10	rbf	0.5	0	25	0.016	0.025
12	GRUPOAVAL	10	rbf	0.5	0	25	0.010	0.020
13	GRUPOSURA	10	rbf	0.5	0	25	0.015	0.028
14	ISA	10	rbf	0.5	0	25	0.014	0.022
15	NUTRESA	10	rbf	0.5	0	25	0.011	0.021
16	PFAVAL	10	rbf	0.5	0	25	0.010	0.016
17	PFCEMARGOS	10	rbf	0.5	0	25	0.012	0.019
18	PFCORFICOL	10	rbf	0.5	0	25	0.008	0.015
19	PFDAVVNDA	10	rbf	0.5	0	25	0.012	0.018
20	PFGRUPOARG	10	rbf	0.5	0	25	0.014	0.024
21	PFGRUPSURA	10	rbf	0.5	0	25	0.013	0.021

Tabla 7 Hiperparámetros y resultados del ajuste del SVR

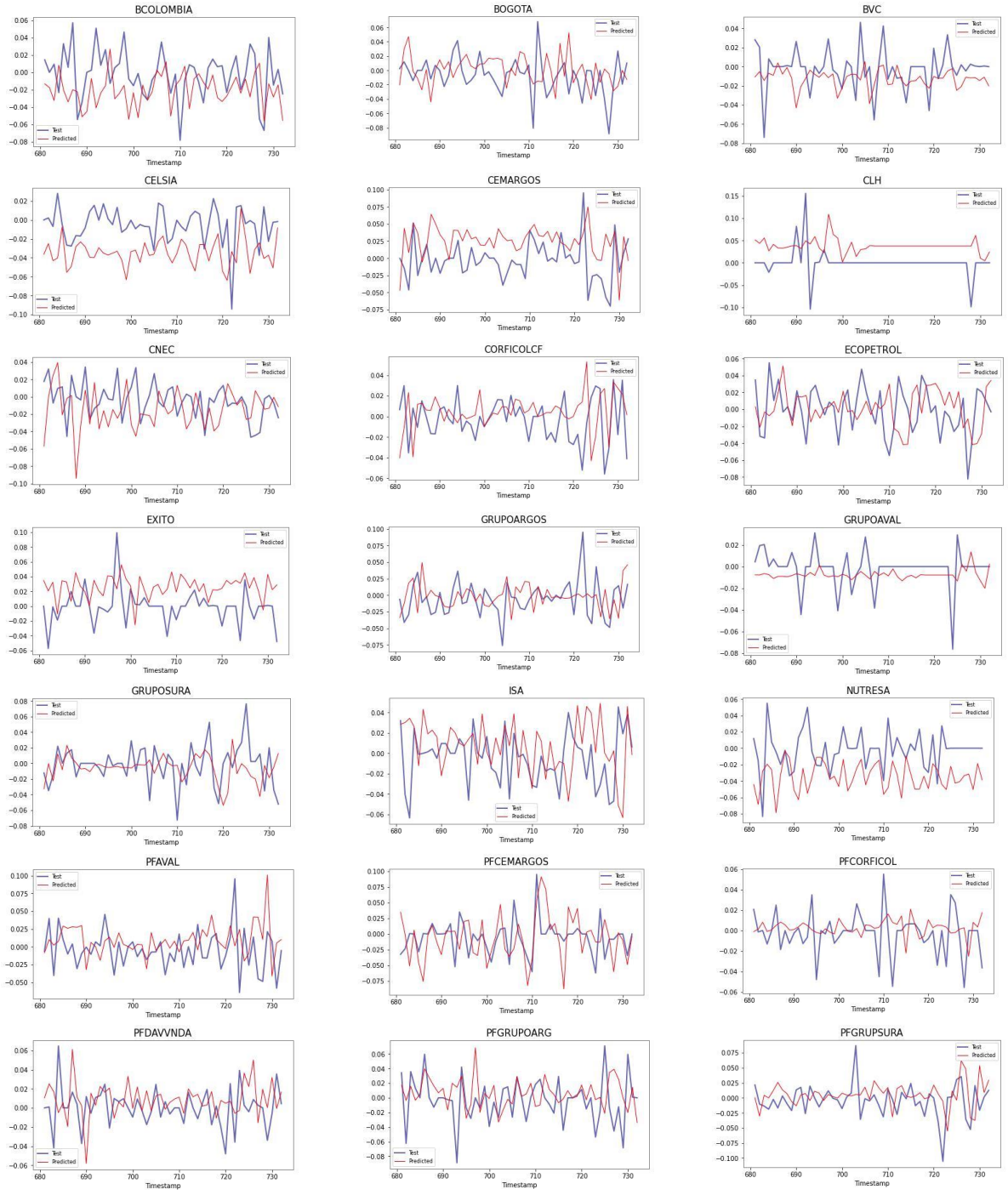


Figura 10 Comparativa proyecciones del SVR vs datos de validación

En la figura 8, se puede notar una comparativa entre la proyección del SVR y los datos de validación. A diferencia de las metodologías estadísticas tradicionales para las series, el SVR parece tener un mejor ajuste. Las acciones que este modelo mejor estaría ajustándose dado que presentan un menor MAE en

comparación con las demás serían Celsia y las acciones de Corficolombiana.

5.2.3 Ajuste del modelo mediante XGboost

En metodologías asociados a bosques aleatorios, es importante el proceso de tunning de hiperparametros, para encontrar aquella combinación que permite que se minimice la métrica de error seleccionada es decir el MAE. A continuación, se presenta el resultado del mejor ajuste, después de un proceso de creación de 218 modelos para cada serie.

índice	Acción	Núm. Rezagos	N estimators	Max depth	Learning rate	Min child weight	RMSE	MAE
1	BCOLOMBIA	10	100	15	0.2	1	0.035	0.028
2	BOGOTA	10	500	15	0.1	1	0.043	0.035
3	BVC	10	100	15	0.2	2	0.022	0.018
4	CELSIA	10	100	15	0.2	2	0.027	0.021
5	CEMARGOS	10	100	15	0.2	2	0.046	0.036
6	CLH	10	100	15	0.1	1	0.049	0.040
7	CNEC	10	100	15	0.2	1	0.025	0.020
8	CORFICOLCF	10	100	15	0.2	1	0.035	0.027
9	ECOPETROL	10	100	15	0.2	2	0.039	0.032
10	EXITO	10	100	15	0.2	2	0.036	0.030
11	GRUPOARGOS	10	100	15	0.2	1	0.035	0.026
12	GRUPOAVAL	5	100	2	0.05	1	0.019	0.014
13	GRUPOSURA	10	100	15	0.2	1	0.028	0.022
14	ISA	10	100	15	0.2	1	0.031	0.023
15	NUTRESA	10	100	15	0.2	2	0.037	0.033
16	PFAVAL	10	100	15	0.2	1	0.037	0.029
17	PFCEMARGOS	10	100	15	0.2	1	0.043	0.035
18	PFCORFICOL	10	500	15	0.2	3	0.024	0.019
19	PFDAVVNDA	10	100	15	0.2	1	0.034	0.028
20	PFGRUPOARG	10	100	15	0.2	3	0.038	0.030
21	PFGRUPSURA	10	100	15	0.2	1	0.036	0.026

Tabla 8 Hiperparámetros y resultados del ajuste del XGboost

En la figura 9, se observa algunos gráficos comparativos de las acciones evaluadas entre el periodo de validación y su respectiva predicción. Además, se adiciona la visualización del pronostico a 90 días con esta metodología, ya que parece que por métricas puede ser la que mejor ofrece resultados para algunas acciones particulares.

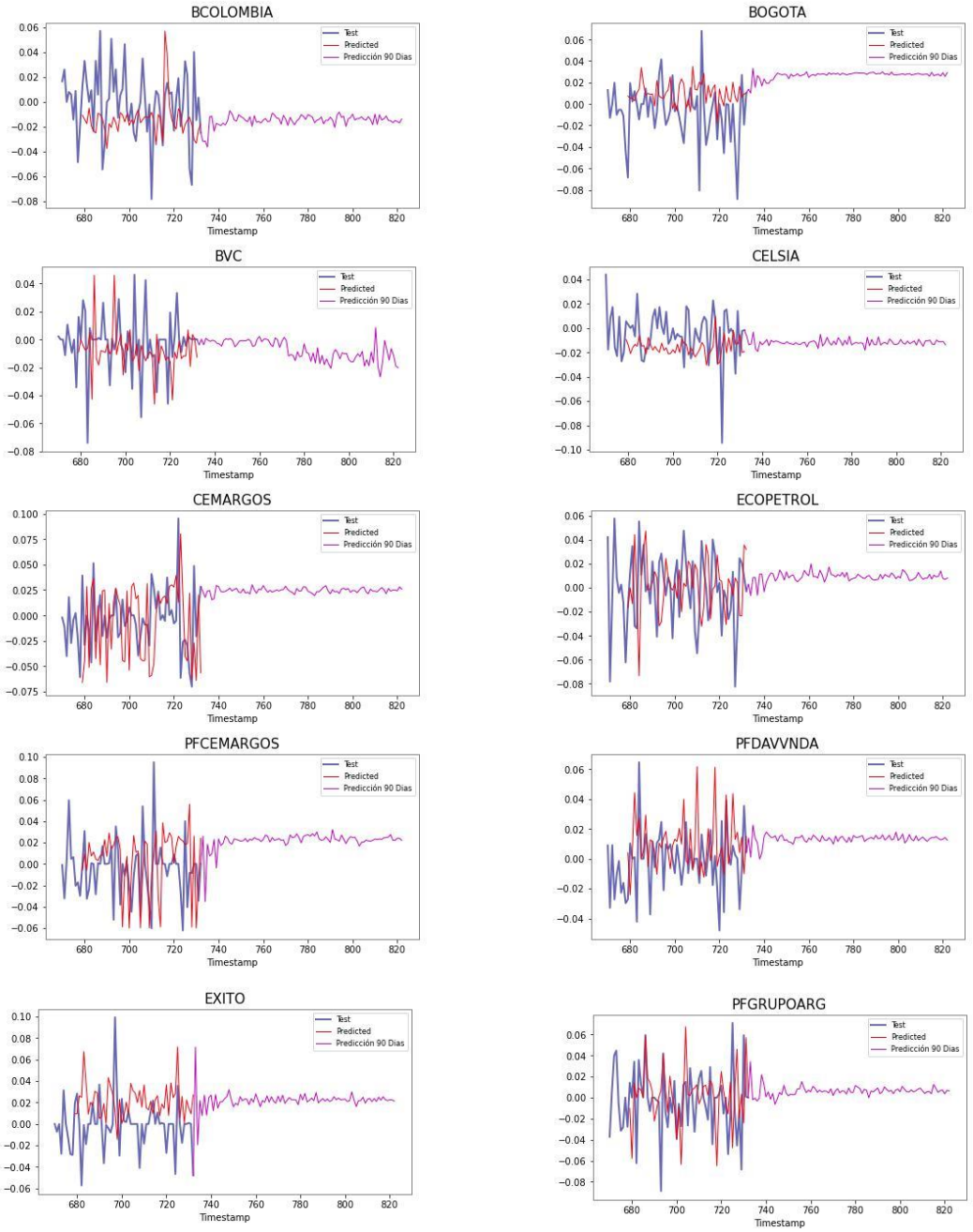


Figura 11 Comparativa proyecciones del XGBoost vs datos de validación y pronóstico (10 acciones)

5.2.4 Ajuste del modelo mediante Redes Neuronales

En redes neuronales es importante contar con modelos que puedan contar con un número importante de épocas, sin necesidad de que eso implique crear modelos que requieran un gasto computacional adicional, al ya presentado con los anteriores modelos construidos. Se obtuvieron las siguientes métricas bajo este enfoque:

Índice	Acción	Epochs	Tamaño Batch	Validation Split	Verbose	RMSE	MAE
1	BCOLOMBIA	500	15	0.1	0	0.023	0.014
2	BOGOTA	500	15	0.1	0	0.029	0.015
3	BVC	500	15	0.1	0	0.015	0.008
4	CELSIA	500	15	0.1	0	0.014	0.008
5	CEMARGOS	500	15	0.1	0	0.023	0.014
6	CLH	500	15	0.1	0	0.024	0.008
7	CNEC	500	15	0.1	0	0.016	0.010
8	CORFICOLCF	500	15	0.1	0	0.017	0.011
9	ECOPETROL	500	15	0.1	0	0.028	0.017
10	EXITO	500	15	0.1	0	0.022	0.010
11	GRUPOARGOS	500	15	0.1	0	0.023	0.011
12	GRUPOAVAL	500	15	0.1	0	0.021	0.009
13	GRUPOSURA	500	15	0.1	0	0.037	0.019
14	ISA	500	15	0.1	0	0.022	0.014
15	NUTRESA	500	15	0.1	0	0.022	0.012
16	PFAVAL	500	15	0.1	0	0.021	0.011
17	PFCEMARGOS	500	15	0.1	0	0.019	0.011
18	PFCORFICOL	500	15	0.1	0	0.016	0.008
19	PFDAVVNDA	500	15	0.1	0	0.016	0.009
20	PFGRUPOARG	500	15	0.1	0	0.021	0.012
21	PFGRUPSURA	500	15	0.1	0	0.019	0.011

Tabla 9 Hiperparametros y resultados del ajuste de Redes Neuronales (LSTM)

En el caso de las redes neuronales, a pesar que en cuanto a métricas se generaron resultados importantes que pueden competir con las métricas obtenidas por metodologías como SVM o XGBoost, el comportamiento del pronóstico a la hora de visualizarlo no genera un mejor resultado al ya obtenido, esto puede deberse a que este tipo de metodologías son débiles a realizar pronósticos en un tiempo como el planteado en este trabajo que es a 90 días.

5.3 Selección del mejor modelo para cada serie

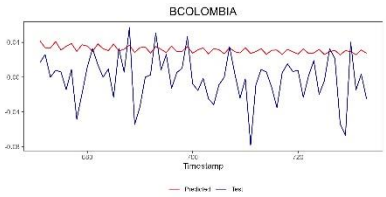
Una vez aplicada cada una de las metodologías escogidas en este proyecto, se seleccionó el modelo que mejor se ajustaba para cada una de las series observadas, quedando de la siguiente forma:

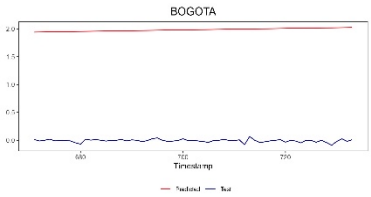
índice	Acción	RMSE	MAE	Modelo
0	BCOLOMBIA	0.0330	0.0261	XGB
1	BOGOTA	0.0332	0.0258	XGB
2	BVC	0.0245	0.0189	SVR
3	CELSIA	0.0228	0.0182	XGB
4	CEMARGOS	0.0449	0.0362	SVR
5	CLH	0.0508	0.0431	SVR
6	CNEC	0.0254	0.0207	XGB
7	CORFICOLCF	0.0305	0.0228	XGB
8	ECOPETROL	0.0369	0.0306	SVR
9	EXITO	0.0372	0.0294	XGB
10	GRUPOARGOS	0.0331	0.0262	SVR
11	GRUPOAVAL	0.0189	0.0137	SVR
12	GRUPOSURA	0.0301	0.0236	SVR
13	ISA	0.0315	0.0244	XGB
14	NUTRESA	0.0389	0.0343	XGB
15	PFAVAL	0.0324	0.0257	XGB
16	PFCEMARGOS	0.0376	0.0304	XGB
17	PFCORFICOL	0.0240	0.0176	SVR
18	PFDAVVNDA	0.0264	0.0207	XGB
19	PFGRUPOARG	0.0387	0.0280	XGB
20	PFGRUPSURA	0.0339	0.0241	XGB

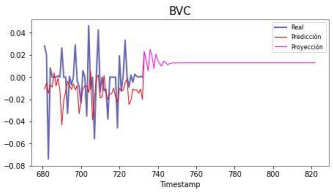
Tabla 10 Modelo escogido para cada serie

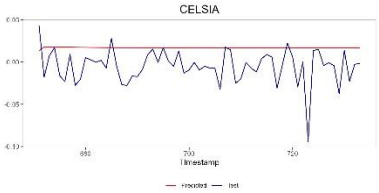
Para hacer esta selección se buscaba priorizar ambas métricas ya que el objetivo independientemente de cómo se midiera, fue contar con el modelo que minimizaba más el error versus otros. Finalmente, la métrica que tuvo mayor relevancia fue el MAE, ya que el objetivo además de contar con un buen modelo, fue contar con una métrica directa del error que además fuese más robusto a la hora de contar eventualmente con valores atípicos.

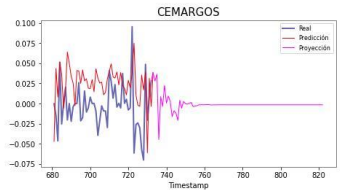
A continuación, se presenta el comportamiento comparativo con la serie histórica, entre los datos reales y las proyecciones.

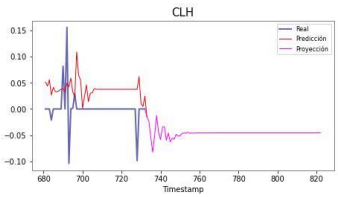

BCOLOMBIA
Timestamp

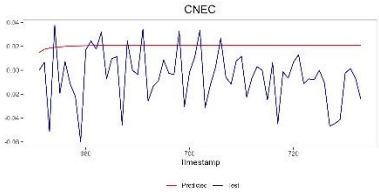

BOGOTA
Timestamp

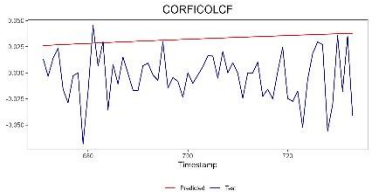

BVC
Real
Predicción
Proyección
Timestamp

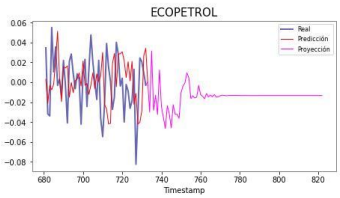

CELSIA
Timestamp

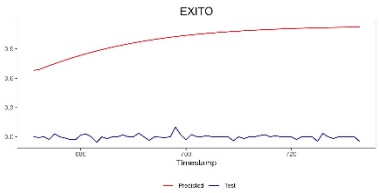

CEMARGOS
Real
Predicción
Proyección
Timestamp

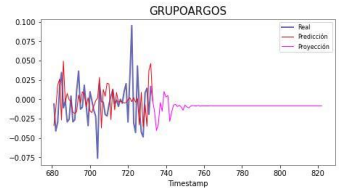

CLH
Real
Predicción
Proyección
Timestamp

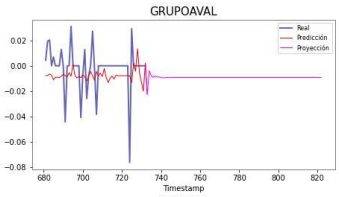

CNEC
Timestamp

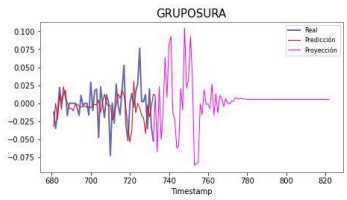

CORFICOLCF
Timestamp

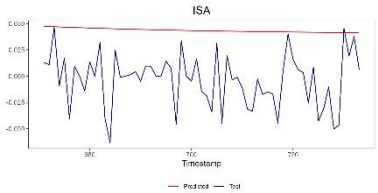

ECOPETROL
Real
Predicción
Proyección
Timestamp

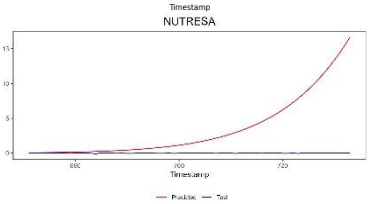

EXITO
Timestamp

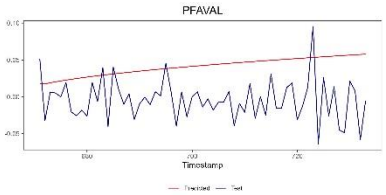

GRUPOARGOS
Real
Predicción
Proyección
Timestamp

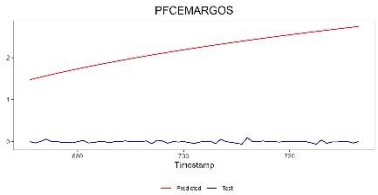

GRUPOAVAL
Real
Predicción
Proyección
Timestamp

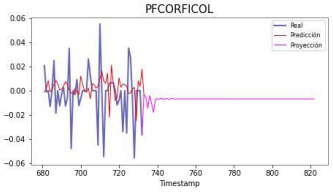

GRUPOSURA
Real
Predicción
Proyección
Timestamp

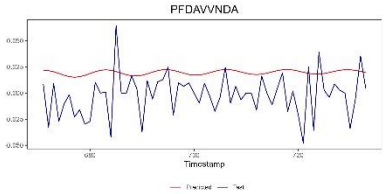

ISA
Timestamp

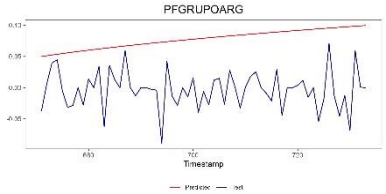

NUTRESA
Timestamp

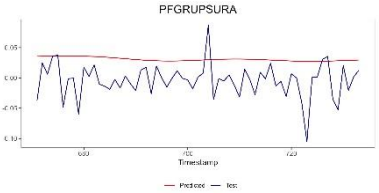

PFAVAL
Timestamp


PFCEMARGOS
Timestamp


PFCORFICOL
Real
Predicción
Proyección
Timestamp


PFDAVVNDA
Timestamp


PFGRUPOARG
Timestamp


PFGRUPSURA
Timestamp

5.4 Definición de portafolio de inversión

Para la definición del portafolio más optimo dado el comportamiento de las series históricas, se decidió trabajar con las acciones que hayan presentado la menor varianza las cuales son:

COMPAÑIA	SIMBOLO
Banco de Bogotá	BOGOTA
Bolsa de Valores de Colombia	BVC
Celsia	CELSIA
Canacol Energy	CNEC
Corficolombiana	CORFICOLCF
Grupo Aval	GRUPOAVAL
Grupo Nutresa	NUTRESA
Grupo Aval	PFAVAL
Corficolombiana	PFCORFICOL
Davivienda	PFDAVVNDA

Tabla 11 Series con la menor varianza en el periodo estudiado

Se decidió para este ejercicio, definir el mejor portafolio posible que se pudiera construir teniendo en cuenta únicamente 5 acciones. Esto nos genera un grupo de 252 combinaciones posible para la definición de ese portafolio.

Una vez calculado la frontera eficiente, los pesos de inversión, el retorno esperado y el sharpe ratio de cada una de las posibles combinaciones que nos arrojan los distintos portafolios que podrían ser factibles, se escoge aquel que presente la mayor ratio sharpe posible, dado que nos dice que es el portafolio que mejor estaría equilibrando la rentabilidad dado el riesgo, dando como resultado el siguiente portafolio optimo.

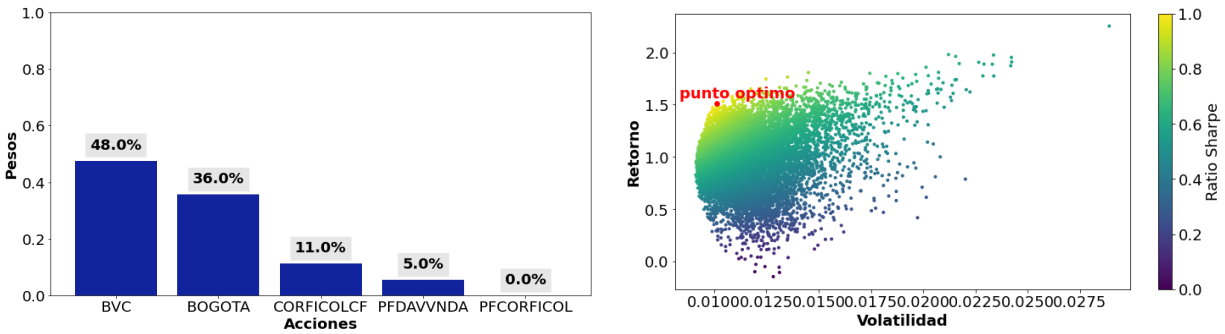


Figura 12 Portafolio optimo conformado por 5 acciones

Es entonces como el portafolio que cumple con el punto óptimo de volatilidad vs retorno, cuenta con un retorno del 153.35%, una volatilidad de 1.03% y un ratio Sharpe de 149.5.

5.5 Construcción del Dashboard en Power BI

Se realiza la construcción de un Dashboard que le permitirá al inversor poder interactuar con la simulación de cada uno de los escenarios que el considere con combinación de las 21 acciones en tamaños de portafolios que el decida. Para esto se usó la herramienta Power Bi, la cual se alimenta con la información histórica de las series y las proyecciones realizadas para cada una de ellas a partir del mejor modelo que se ajustó a ella. Este consta de 3 páginas, donde en la primera veremos gráficamente las descriptivas de la serie que nos permita conocer y visualizar el comportamiento histórico de la misma.

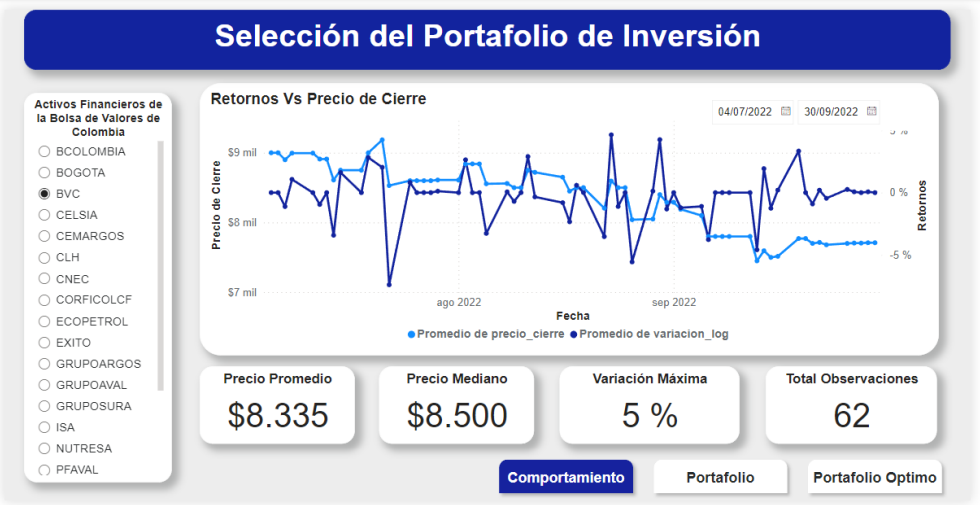


Figura 13 Pagina 1 Dashboard: Análisis descriptivo de la serie

En la página 2, se encuentra el cálculo de la frontera eficiente y la distribución de la inversión de mínimo riesgo basada en el conjunto de acciones que el inversor desee escoger, este proceso es iterativo, y se puede seleccionar el tamaño del portafolio que se desee. Se le entregará al inversor la distribución de la inversión, la rentabilidad esperada, la volatilidad esperada y el Ratio Sharpe asociado al portafolio escogido.

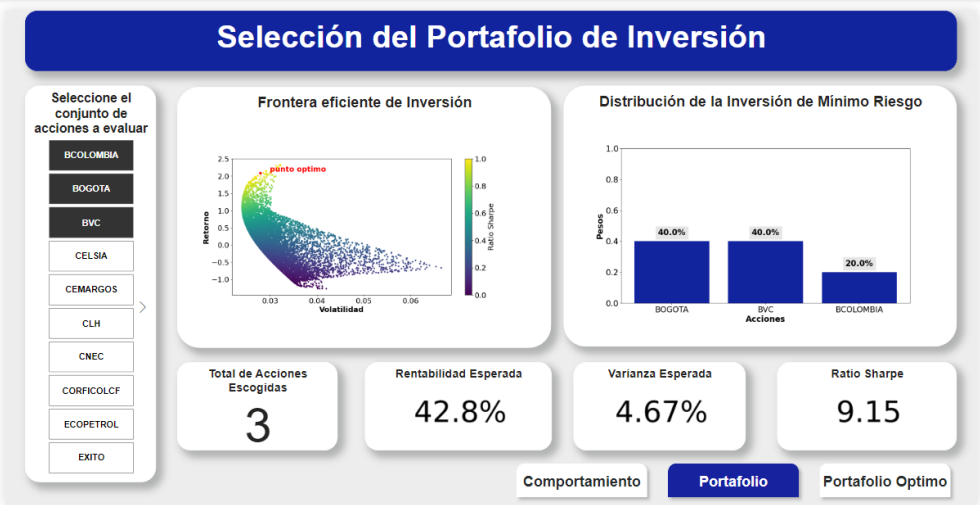


Figura 14 Pagina 2 Dashboard: Definición de la rentabilidad esperada del portafolio escogido por el inversor

En la última página del Dashboard, se le entrega al inversor el portafolio óptimo de inversión dado el tamaño que este escoja. Se le entrega tanto la rentabilidad esperada, como la visualización de la frontera eficiente y la distribución de la inversión a realizar.

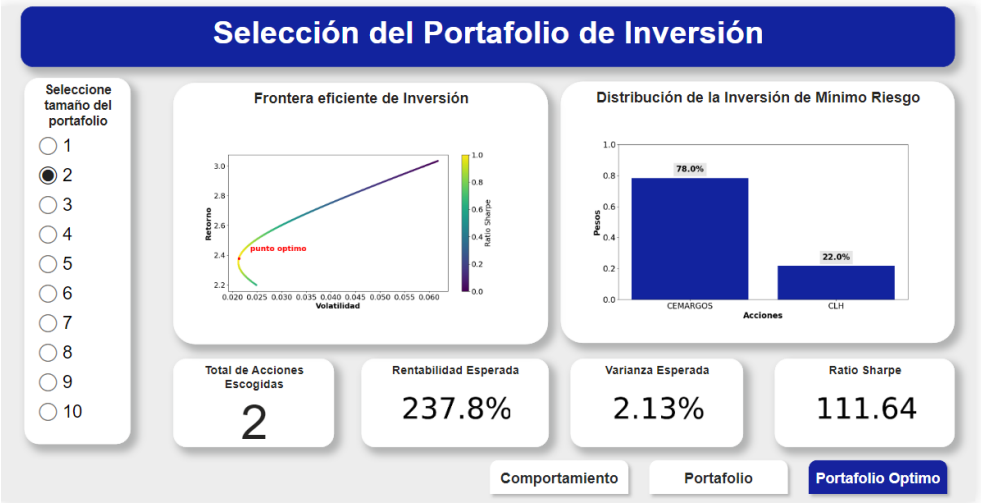


Figura 15 Pagina 3 Dashboard: Definición del portafolio optimo dado un tamaño de este

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. CONCLUSIONES

El uso de metodologías de Machine Learning para la estimación de la volatilidad de activos financieros puede mejorar la precisión de las predicciones en comparación con las metodologías tradicionales, sin embargo, se debe considerar el adecuado uso de los métodos y la preparación de los datos debe responder de manera adecuada, para garantizar robustes en los resultados y principalmente si tenemos en cuenta el objetivo de reducir la incertidumbre en la toma de decisiones para un inversionista.

La selección de activos financieros bajo criterio de mínima varianza permite incorporar criterios de negocio y de mercados adecuados para el contexto, tales como una adecuada diversificación, para reducir el riesgo de invertir en activos relacionados con una misma empresa o grupo económico. Adicionalmente genera beneficios computacionales, logrando un set de iteraciones de menor dimensión y conseguidas en menor tiempo, sin sacrificar la premisa principal de este trabajo que es encontrar una diversificación eficiente para un inversionista.

La estimación del mejor portafolio utilizando la Teoría de Markowitz pueden ayudar a minimizar el riesgo de inversión y maximizar la rentabilidad esperada, esto mediante la evaluación de 252 portafolios factibles que les permite a los inversores seleccionar una combinación adecuada de activos financieros en función de su nivel de riesgo y tolerancia al mismo.

La utilización del criterio de Sharpe en la evaluación de la rentabilidad del portafolio permite comparar el rendimiento obtenido con la cantidad de riesgo asumido y, por lo tanto, determinar la eficiencia del portafolio, la metodología propuesta puede ser una herramienta valiosa para los inversores y analistas financieros en la toma de decisiones de inversión y en la gestión de riesgos.

La Teoría de Markowitz es una herramienta eficaz para la construcción de portafolios diversificados y bien equilibrados que minimicen el riesgo y maximicen la rentabilidad, de igual manera la selección de activos financieros adecuados es fundamental para lograr un rendimiento satisfactorio y minimizar el riesgo.

La combinación de enfoques tradicionales y de aprendizaje automático puede mejorar la precisión de las predicciones y ayudar a los inversores a tomar decisiones más informadas, en horizontes de tiempo más amplios y lograr tener una mejor gestión de sus rendimientos y exposición al riesgo.

La construcción de un portafolio diversificado es una estrategia efectiva para reducir el riesgo y maximizar la rentabilidad esperada, La metodología propuesta en esta tesis puede ser una guía útil para los inversores en la selección de activos financieros y en la construcción de portafolios bien equilibrados y diversificados.

6.2. TRABAJOS FUTUROS

Basado en las conclusiones del desarrollo de este trabajo, se pueden plantear varios trabajos futuros y otras aplicaciones de las metodologías utilizadas en la investigación:

- Ampliar el análisis de la volatilidad de activos financieros a otras bolsas de valores: Se pueden aplicar las metodologías de Machine Learning y la Teoría de Markowitz a otras bolsas de valores, con el objetivo de identificar patrones y relaciones similares y poder realizar comparaciones Inter bolsas.
- Análisis de los factores macroeconómicos que influyen en la volatilidad: Se pueden investigar los factores macroeconómicos que influyen en la volatilidad de los activos financieros y cómo se pueden incluir en los modelos de Machine Learning y de la Teoría de Markowitz para mejorar la precisión de la estimación de la volatilidad.
- Desarrollo de nuevas técnicas de selección de activos financieros: Se pueden explorar nuevas técnicas para la selección de los activos financieros en la construcción del portafolio, como la utilización de técnicas de clustering o de reducción de dimensionalidad para identificar grupos de activos similares o correlacionados.
- Incorporación de otras fuentes de datos: Se pueden explorar otras fuentes de datos, como datos no estructurados (por ejemplo, noticias, redes sociales) o información de transacciones de alta frecuencia, para mejorar la precisión de las estimaciones de la volatilidad y de la selección de los activos.
- Análisis de la influencia del tamaño de la muestra en la precisión de las estimaciones: Se puede realizar un análisis para determinar cómo el tamaño de la muestra utilizado para el análisis de la volatilidad y la selección de los activos financieros influye en la precisión de las estimaciones y cómo se pueden optimizar las muestras.
- Evaluación de la influencia de la estrategia de reequilibrio del portafolio: Se puede investigar cómo la frecuencia y la estrategia de reequilibrio del portafolio influyen en la rentabilidad y en la reducción del riesgo del portafolio.
- Aplicaciones en la gestión de riesgos: Se pueden explorar otras aplicaciones de las metodologías de Machine Learning y de la Teoría de Markowitz en la gestión de riesgos financieros, como la identificación de eventos extremos y la elaboración de planes de contingencia y gestión de crisis.

ANEXOS

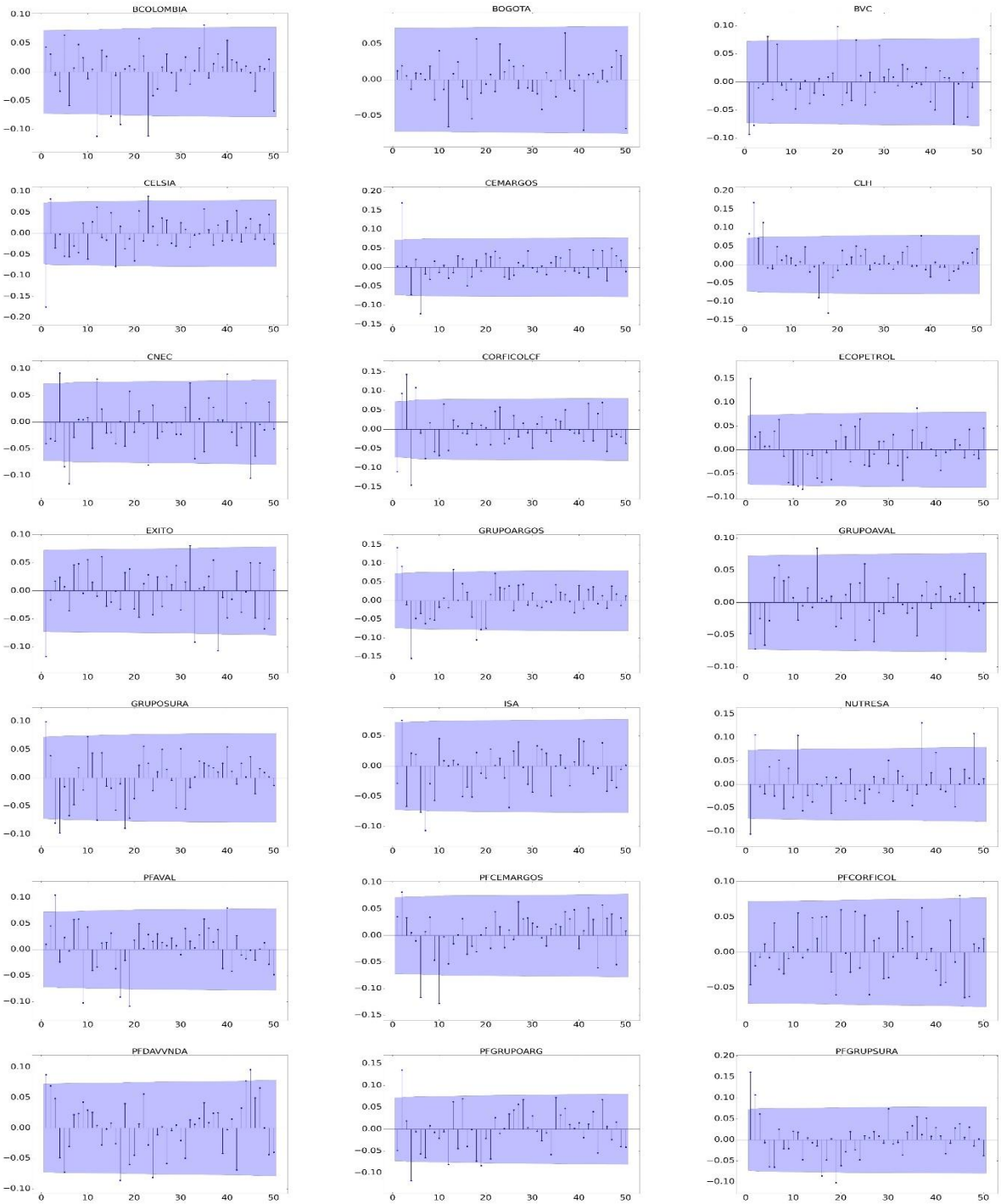


Figura: Grafico del FAC para las series de la BVC consideradas

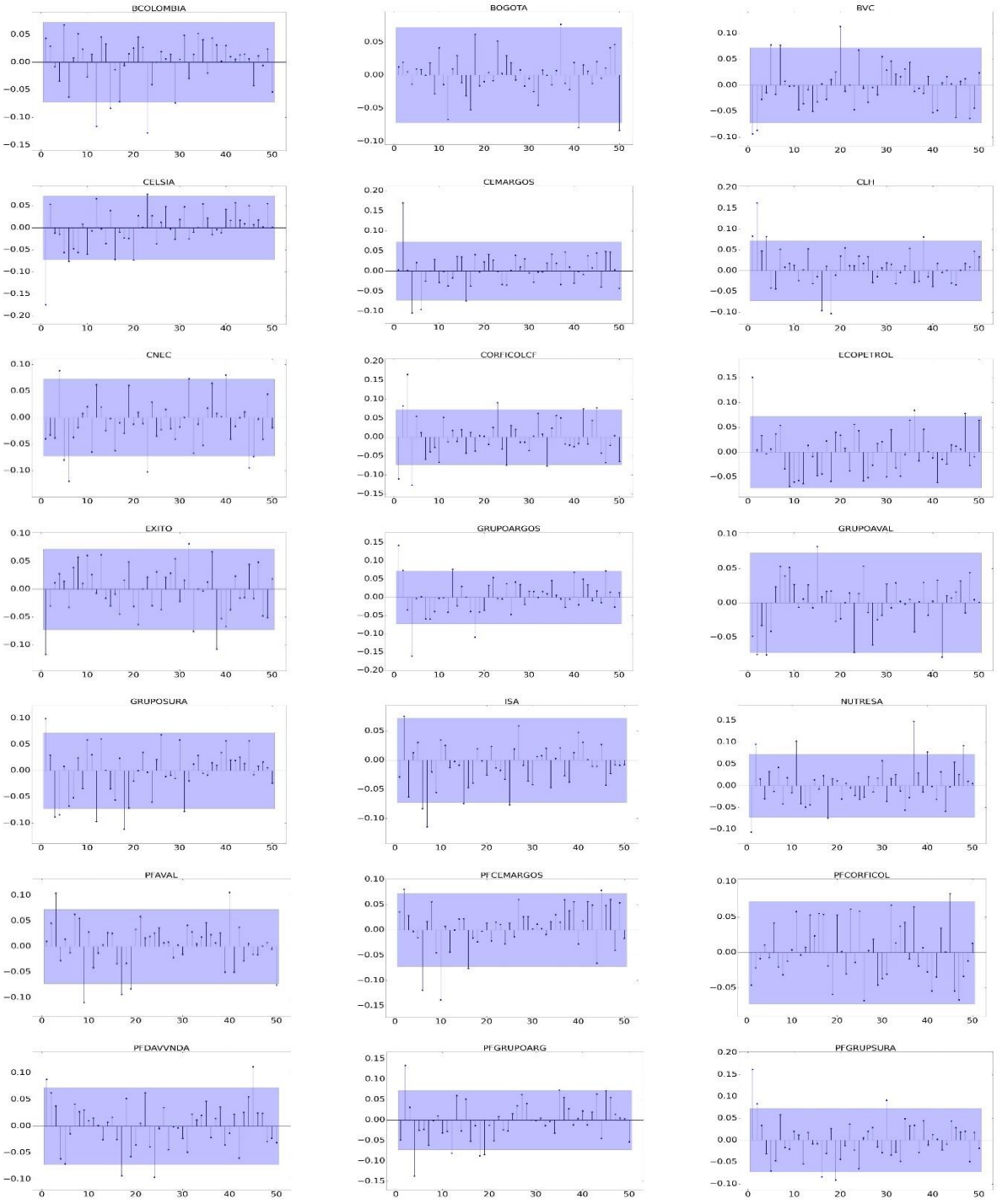


Figura: Grafico de PCAF para las series de la BVC consideradas

índice	Acción	ADF Statistic	P Value	Resultado Hipótesis
0	BCOLOMBIA	-25,76	0,00%	Estacionaria
1	BOGOTA	-26,58	0,00%	Estacionaria
2	BVC	-9,49	0,00%	Estacionaria
3	CELSIA	-19,56	0,00%	Estacionaria
4	CEMARGOS	-12,10	0,00%	Estacionaria
5	CLH	-10,67	0,00%	Estacionaria
6	CNEC	-12,90	0,00%	Estacionaria
7	CORFICOLCF	-11,22	0,00%	Estacionaria
8	ECOPETROL	-23,10	0,00%	Estacionaria
9	EXITO	-30,38	0,00%	Estacionaria
10	GRUPOARGOS	-15,07	0,00%	Estacionaria
11	GRUPOAVAL	-15,61	0,00%	Estacionaria
12	GRUPOSURA	-7,82	0,00%	Estacionaria
13	ISA	-12,09	0,00%	Estacionaria
14	NUTRESA	-18,14	0,00%	Estacionaria
15	PFAVAL	-8,65	0,00%	Estacionaria
16	PFCEMARGOS	-10,10	0,00%	Estacionaria
17	PFCORFICOL	-28,27	0,00%	Estacionaria
18	PFDAVVNDA	-12,75	0,00%	Estacionaria
19	PFGRUPOARG	-14,34	0,00%	Estacionaria
20	PFGRUPSURA	-12,21	0,00%	Estacionaria

Tabla: Resultado de la prueba Dickey Fuller

7. REFERENCIAS BIBLIOGRÁFICAS

[1] A. P. A. R. L. &. C. Á. Calvo, “Manual del sistema financiero español,” 2018.

[2] H. Escobar and V. Cuartas, “Diccionario Economico Financiero,” *Invest*, 2016.

[3] J. Crespo and C. Fernandez, “Descubriendo el Mercado (I): medidas de posición central,” *Bolsa: revista mensual de bolsas y mercados españoles*, vol. I, no. 178, pp. 61–64, 2008.

[4] I. Palacio Bascon and Rodriguez Muñoz, “La bolsa de valores. El diario de un inversor medio,” *Universidad de Cordoba*, 2020.

[5] Castillo Garavito and Camacho Padilla, “Fundamentos de la creación y cálculo de los indices Colcap y Col 20 y el mercado de derivados en Colombia,” *Universidad Autonoma de Bucaramanga UNAB*, 2008.

[6] C. Hayashi, *What is data science? Fundamental concepts and a heuristic example.* Tokyo: Springer, 1998.

[7] T. Grau, “Forecasting Asset Portfolio Market Risk,” 2004.

[8] A. Novales, *Modelos ARCH univariantes y multivariantes*. Madrid, España: Universidad Complutense de Madrid, 2013.

[9] Guerra Rodriguez, “Fundamentos y variantes de los modelos ARIMA para el analisis de series temporales,” *Aplicación a la estadística universitaria*, 2022.

[10] J. Alzubi, N. Anand, and K. Akshi, “Machine learning from theory to algorithms: an overview,” *J Phys Conf Ser*, vol. 1142, no. 1, 2018.

[11] W. S. Noble, “What is a support vector machine?” *Nat Biotechnol*, vol. 24, no. 12, pp. 1565–1567, 2006.

[12] L. Gui, “Financial time series forecasting using support vector machine,” *Tenth International Conference on Computational Intelligence and Security*, pp. 39–43, 2014.

[13] L. Cao and H. Tay, “Support vector machine with adaptive parameters in financial time series forecasting,” *IEEE Trans Neural Netw*, vol. 14, no. 6, pp. 1506–1518, 2003.

[14] I. Paliari, A. Karanikola, and S. Kotsaintis, “A comparison of the optimized LSTM, XGBOOST and ARIMA in Time Series forecasting,” in *12th International Coference on Information, Intelligence, Systems applications (IISA)*, Chania Crete, Greece, 2021.

[15] S. Frutos Serrano, “Comparación entre el XGBoost y Regresión Lineal Múltiple para la predicción de la evolución del precio de las acciones,” 2022.

[16] J. Connor, D. Martin, and L. Atlas, “Recurrent neural networks and robust time series prediction,” *IEEE Trans Neural Netw*, vol. 5, no. 2, pp. 240–254, 1994.

[17] A. Yadav, C. Jha, and A. Sharan, “Optimizing LSTM for time series prediction in Indian stock market,” *Procedia Comput Sci*, 2020.

[18] Yamak, Yujian, and Gadosey, “A comparison between arima, lstm, and gru for time series forecasting,” *Proceedings of the 2019 2nd International Conference on Algorithms*, pp. 49–55, 2019.

[19] Moscote Floréz, *Elementos de estadística en riesgo financiero*. USTA, 2013.

[20] Willmott and Matsuura, “Ambiguities inherent in sums-of-squares-based error statistics,”

Ambiguities inherent in sums-of-squares-based error statistics, 2009.
[21] Guerra Espinoza, “Predicción de precios de cierre de acciones en la bolsa de valores mediante técnicas de aprendizaje automático y minería de datos: evaluación de escenarios para transacciones intradía,” 2021.
[22] Delgado Vélez and Durango Gutierréz, “Estructuración de un portafolio de inversiones con acciones colombianas,” *Semestre Económico*, vol. 21, no. 46, pp. 167–183, 2018.
[23] Cuartas Arias, “Optimización financiera de portafolios de inversión en la Bolsa de Valores de Colombia mediante el uso de RGOP,” Universidad EAFIT, Medellin, 2016.
[24] R. Wild, “Index Investing for Dummies,” *John Wiley & Sons*, 2008.
[25] Z. Zhang, S. Zohren, and S. Roberts, “Deep learning for portfolio optimization,” *The Journal of Financial Data Science*, vol. 2, no. 4, pp. 8–20, 2020.
[26] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. 2021.
[27] F. Garza, “Forecasting with Machine Learning Models,” *https://towardsdatascience.com/forecasting-with-machine-learning-models-95a6b6579090*, 2021.
[28] J. Amat Rodrigo, “ Skforecast Docs 0.5.1,” *https://joaquinamatrodrigo.github.io/skforecast/0.5.1/index.html*, 2022.