UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Estrategias de Creación de Carteras de Inversión Basadas en Ciencia de Datos

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Llobregat Ruiz, Pablo

Tutor/a: Jordán Prunera, Jaume Magí

Director/a Experimental: Martínez Barbero, Xavier

CURSO ACADÉMICO: 2023/2024

Resum

Aquest treball final de grau explora l'aplicació de la ciència de dades en l'elaboració d'estratègies d'inversió, utilitzant models predictius avançats per a millorar el rendiment financer. Es desenvolupen i validen diversos models, incloent regressió lineal, xarxes neuronals, i mètodes d'ensemble, demostrant la seva eficàcia en superar els índexs de referència com l'S&P 500. Els resultats destacats suggereixen que l'adopció d'aquestes tècniques pot proporcionar avantatges significatius en la gestió de carteres d'inversió.

Paraules clau: ciència de dades, estratègies d'inversió, rendiment financer, models predictius, anàlisi quantitativa

Resumen

Este trabajo de fin de grado investiga la aplicación de la ciencia de datos en la creación de estrategias de inversión, empleando modelos predictivos avanzados para mejorar el rendimiento financiero. Se desarrollan y validan varios modelos, incluyendo regresión lineal, redes neuronales y métodos de ensamble, mostrando su efectividad en superar índices de referencia como el S&P 500. Los resultados indican que la integración de estas técnicas puede ofrecer ventajas significativas en la gestión de carteras de inversión.

Palabras clave: ciencia de datos, estrategias de inversión, rendimiento financiero, modelos predictivos, análisis cuantitativo

Abstract

This undergraduate thesis examines the application of data science in developing investment strategies, utilizing advanced predictive models to enhance financial performance. Various models are developed and validated, including linear regression, neural networks, and ensemble methods, demonstrating their effectiveness in outperforming benchmark indices such as the S&P 500. The findings suggest that adopting these techniques can provide significant benefits in investment portfolio management.

Key words: data science, investment strategies, financial performance, predictive models, quantitative analysis

Índice general

Índice de figuras

Índice de tablas

Glosario

Términos Generales

API **Application Programming Interface**. Interfaz de programación de aplicaciones que permite la comunicación entre diferentes sistemas de software.

CSV **Comma-Separated Values**. Formato de archivo utilizado para almacenar datos tabulares en texto plano, donde los valores están separados por comas.

EVT **Extreme Value Theory**. Teoría del valor extremo, utilizada en estadística para modelar eventos extremos y raros.

MAE **Mean Absolute Error**. Error absoluto medio, una métrica utilizada para evaluar la precisión de un modelo predictivo.

MSE **Mean Squared Error**. Error cuadrático medio, una métrica que mide la diferencia promedio al cuadrado entre los valores observados y los valores predichos por un modelo.

ODS **Objetivos de Desarrollo Sostenible**. Conjunto de objetivos globales establecidos por las Naciones Unidas para abordar desafíos globales, incluyendo pobreza, desigualdad y cambio climático.

RF **Random Forest**. Bosque aleatorio, un algoritmo de aprendizaje automático que consiste en múltiples árboles de decisión.

R^2 **Coeficiente de Determinación**. Medida estadística que indica la proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes.

TFG **Trabajo de Fin de Grado**. Proyecto final que los estudiantes deben completar para obtener su título de grado.

VAR **Value at Risk**. Valor en riesgo, una medida de riesgo financiero que estima la cantidad máxima que se podría perder en una cartera de inversión en un período determinado con un nivel de confianza específico.

XGBoost **Extreme Gradient Boosting**. Algoritmo de aprendizaje automático basado en técnicas de boosting que es eficaz para tareas de clasificación y regresión.

Términos Financieros

ROA **Return on Assets**. Retorno sobre activos, una medida de rentabilidad que indica cuán eficiente es una empresa en generar ganancias a partir de sus activos totales.

ROE Return on Equity. Retorno sobre el patrimonio, una medida de rentabilidad que indica cuán eficiente es una empresa en generar ganancias a partir del capital de los accionistas.

ROIC Return on Invested Capital. Retorno sobre el capital invertido, una medida de rentabilidad que indica la eficiencia de una empresa en utilizar el capital invertido en sus operaciones para generar ganancias.

BVPS Book Value Per Share. Valor contable por acción, una medida del valor contable de una empresa en función de cada acción en circulación.

TSR Total Shareholder Return. Retorno total para el accionista, una medida del rendimiento total de una inversión en una acción, incluyendo la apreciación del precio y los dividendos.

PE Ratio Price-Earnings Ratio. Relación precio-ganancias, una medida de valoración que compara el precio de la acción de una empresa con sus ganancias por acción.

EBITDA Earnings Before Interest, Taxes, Depreciation, and Amortization. Beneficios antes de intereses, impuestos, depreciación y amortización, una medida de la rentabilidad operativa de una empresa.

EPS Earnings Per Share. Beneficio por acción, una medida de la rentabilidad de una empresa calculada como las ganancias netas divididas por el número de acciones en circulación.

Términos Estadísticos y de Modelos

ARD Automatic Relevance Determination. Técnica de regresión que determina automáticamente la relevancia de las variables de entrada.

MLP Multilayer Perceptron. Perceptrón multicapa, un tipo de red neuronal utilizada en aprendizaje automático.

Índices y Mercados

SP 500 Standard Poor's 500. Índice bursátil que incluye las 500 empresas más grandes que cotizan en bolsas de valores en los Estados Unidos.

ETF Exchange-Traded Fund. Fondo cotizado en bolsa, un tipo de inversión que se negocia en bolsas de valores como una acción.

CAPÍTULO 1
Introducción

"Si no analizas las empresas, tienes las mismas oportunidades de éxito que un jugador de póker apostando sin mirar las cartas" – Peter Lynch

1.1 Motivación y Propósito del Estudio

El mundo de las inversiones ha cambiado radicalmente con la llegada de la era digital. La abundancia de datos y el desarrollo de nuevas tecnologías han transformado el análisis financiero. Los inversores buscan constantemente superar los índices de referencia como el S&P 500. Esto plantea un desafío y una oportunidad.

La **ciencia de datos** emerge como una herramienta revolucionaria en este contexto. Permite descubrir patrones ocultos en grandes volúmenes de datos. También facilita la predicción de tendencias del mercado. Estas capacidades pueden llevar a la creación de estrategias de inversión más efectivas [2].

Este estudio surge de la **necesidad** de explorar hasta qué punto las técnicas avanzadas de ciencia de datos pueden mejorar las estrategias de inversión tradicionales. Específicamente, busca determinar si estos métodos pueden ofrecer una ventaja competitiva sustancial y sostenible.

El **propósito de este Trabajo Fin de Grado (TFG)** es doble. Primero, implementar y evaluar varios modelos predictivos, como el **aprendizaje automático** y el análisis cuantitativo. Segundo, verificar si estos modelos pueden efectivamente superar el rendimiento del S&P 500 al ajustar por volatilidad y otros factores de riesgo.

Además, este trabajo pretende llenar un vacío en la literatura existente. Aunque hay estudios sobre la aplicación de la ciencia de datos en las finanzas [10], pocos han abordado su impacto real sobre el rendimiento de las inversiones en comparación con índices estándar.

En resumen, la investigación busca validar la hipótesis de que las técnicas de ciencia de datos no solo son viables sino también preferibles en la creación de carteras (o portafolios) de inversión. Se espera que los hallazgos contribuyan significativamente a la práctica y teoría financiera. También se pretende que proporcionen una guía práctica para los inversores que desean incorporar métodos científicos en la gestión de sus carteras.

1.2 Objetivos de la Investigación

El enfoque principal de esta investigación es evaluar la eficacia de las técnicas de ciencia de datos en la optimización de estrategias de creación de carteras de inversión. A continuación, se detallan los objetivos específicos que guían este estudio:

1. **Obtención y Preparación de Datos:**

   Recopilar, limpiar y preparar datos financieros relevantes para su uso en modelos predictivos. Este objetivo busca asegurar la calidad y coherencia de los datos utilizados en el estudio. Se incluirán procesos de scraping, integración de múltiples fuentes de datos, manejo de datos faltantes y normalización de datos.

2. **Desarrollar Modelos Predictivos Eficientes:**

   Implementar diversos modelos de aprendizaje automático y análisis cuantitativo para predecir tendencias del mercado financiero. La meta es identificar aquellos modelos que mejor se adaptan a la volatilidad y dinámica del mercado. Se explorarán modelos como la regresión lineal, regresión Ridge, Lasso, ARD, Random Forest, XGBoost y redes neuronales, evaluando su capacidad para captar patrones en datos históricos y su rendimiento predictivo.

3. **Optimización de la Selección de Características:**

   Investigar y aplicar técnicas avanzadas para la selección de características que contribuyan significativamente al rendimiento predictivo de los modelos. Este objetivo busca maximizar la eficiencia de los modelos al reducir la dimensionalidad de los datos. Se utilizarán métodos como la selección basada en importancia de características, técnicas de reducción de dimensionalidad como PCA, y enfoques de selección automática como Lasso y ARD.

4. **Comparar el Rendimiento contra el S&P 500:**

   Evaluar si los modelos desarrollados logran superar el rendimiento del índice S&P 500 en términos de retorno ajustado por riesgo. Esto incluirá análisis de rentabilidad y comparaciones de volatilidad. Se realizará una evaluación detallada de los resultados obtenidos mediante métricas de rendimiento financiero como el retorno anualizado, la relación Sharpe y el tracking error.

5. **Análisis de Riesgo y Volatilidad:**

   Examinar cómo las estrategias basadas en ciencia de datos afectan los factores de riesgo asociados con la inversión. Esto ayudará a entender mejor las compensaciones entre riesgo y retorno en contextos de inversión optimizados por ciencia de datos. Se analizarán medidas de riesgo como la desviación estándar, el Value at Risk (VaR) y el Conditional Value at Risk (CVaR), y se propondrán estrategias de mitigación de riesgo basadas en técnicas de cobertura y diversificación.

6. **Contribuciones a la Teoría Financiera:**

   Contribuir a la literatura académica mediante el análisis y discusión de los resultados obtenidos. Se espera que este estudio ofrezca nuevas perspectivas sobre la utilización de la ciencia de datos en la gestión financiera. Se documentarán las metodologías aplicadas, los hallazgos clave y las implicaciones prácticas para inversores y gestores de carteras, destacando cómo las técnicas de ciencia de datos pueden integrarse en la toma de decisiones financieras para mejorar el rendimiento de las inversiones.

CAPÍTULO 2
Fundamentos Teóricos

2.1 Teoría de Carteras y Diversificación

La teoría de carteras moderna, originada por Harry Markowitz [13], revolucionó nuestra comprensión sobre la diversificación de inversiones. Esta teoría argumenta que no solo es importante la selección de activos de alto rendimiento, sino también cómo estos activos se combinan.

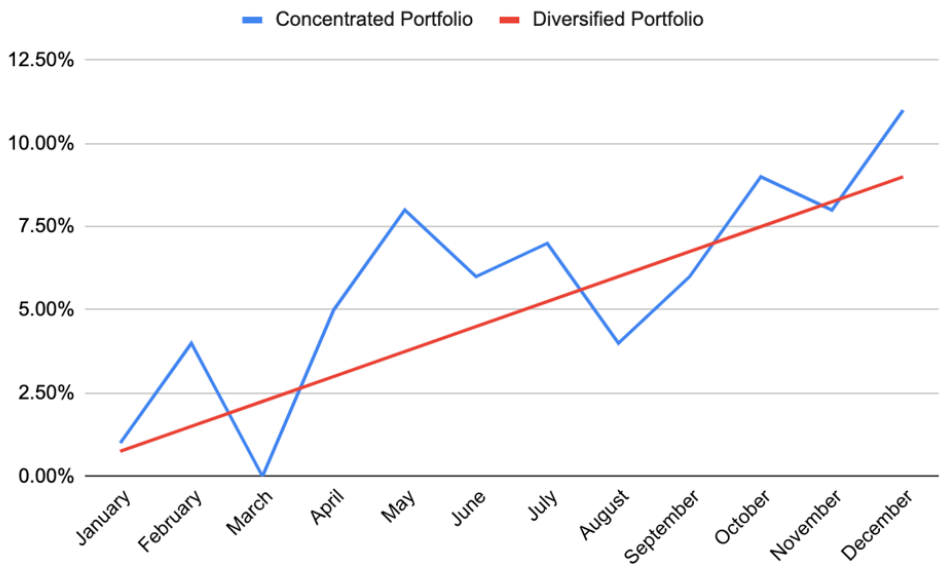

Figura 2.1: Cartera Concentrada vs Diversificada

La premisa es simple: la diversificación reduce el riesgo no sistemático mediante la combinación de activos no correlacionados [25]. Si un activo no proporciona resultados buenos, otro puede prosperar, equilibrando las pérdidas y reduciendo la volatilidad general de la cartera. Además, la relación entre riesgo y retorno es fundamental en la teoría de carteras. Generalmente, un mayor riesgo implica un mayor retorno potencial. Este trade-off entre riesgo y retorno es un concepto clave que debe ser considerado al optimizar una cartera de inversión.

La optimización de carteras se basa en gran medida en la estimación precisa de los retornos esperados y la matriz de varianza-covarianza. Sin embargo, la sensibilidad de las carteras eficientes a los cambios en estas estimaciones puede llevar a un rendimiento subóptimo fuera de muestra. Para mitigar estos problemas, se han desarrollado méto-

dos como el modelo de Black-Litterman [29], estimadores bayesianos [3], y estimadores robustos [15].

En años recientes, la inteligencia artificial ha demostrado su capacidad para mejorar la estimación de retornos esperados, empleando técnicas de aprendizaje automático para lograr una mayor precisión en la predicción de retornos y covarianzas [11]. Concretamente, las redes neuronales de memoria a largo corto plazo (LSTM) han mostrado ser efectivas en la predicción de precios de acciones y la creación de carteras basadas en estas predicciones.

2.2 Optimización de Carteras Basada en Predicciones

La optimización de carteras basada en predicciones busca superar las limitaciones de los modelos tradicionales de media-varianza, los cuales dependen en gran medida de la precisión de los parámetros de entrada. Las redes LSTM, en particular, han demostrado ser eficaces en la mejora de estas predicciones al capturar dependencias a largo plazo y relaciones no lineales en los datos financieros.

Xavier Martínez-Barbero et al. [22] combinan técnicas clásicas de optimización de media-varianza con redes neuronales LSTM para proporcionar predicciones de retornos más precisas y generar carteras rentables para varios periodos de inversión. Su investigación muestra que las carteras basadas en predicciones consistentes superan al índice EURO STOXX 50, incluso en mercados bajistas, logrando menores errores predictivos y mayores retornos ajustados por riesgo.

2.3 Literatura Actual y Contribuciones

Desde la introducción del modelo de selección de carteras de media-varianza por Harry Markowitz en 1952 [25], se han aplicado diversos enfoques para abordar la optimización de carteras. Sin embargo, uno de los principales desafíos sigue siendo la estimación precisa de los parámetros de entrada. Estudios recientes [22] han demostrado que los algoritmos de aprendizaje automático, como las redes LSTM, pueden mejorar significativamente estas estimaciones, proporcionando carteras más robustas y capaces de generar retornos consistentes.

Este trabajo se diferencia de la literatura existente al evaluar la eficacia de los modelos predictivos en diferentes contextos de mercado, tanto alcistas como bajistas, y al utilizar datos históricos para entrenar y validar los modelos, demostrando su capacidad para generalizar y predecir con precisión en diferentes entornos financieros.

La premisa es simple. La diversificación reduce el riesgo no sistemático mediante la combinación de activos no correlacionados [25]. Si un activo falla, otro puede prosperar, equilibrando las pérdidas y manteniendo la rentabilidad general de la cartera.

Es importante entender las diferencias entre el riesgo sistemático y el riesgo no sistemático. El riesgo sistemático, también conocido como riesgo de mercado, es el riesgo inherente a todo el mercado o segmento del mercado. Este tipo de riesgo es causado por factores externos como cambios en las tasas de interés, inflación, recesiones económicas o eventos políticos. Debido a su naturaleza, el riesgo sistemático no puede ser eliminado mediante la diversificación, ya que afecta a todas las inversiones dentro de un mercado. Por ejemplo, una crisis económica global afectará negativamente a la mayoría de los activos en el mercado, independientemente de cuán diversificada esté la cartera.

Por otro lado, el riesgo no sistemático, también conocido como riesgo específico o idiosincrático, es el riesgo asociado a una empresa o industria específica. Este tipo de riesgo puede ser reducido o eliminado mediante la diversificación, ya que no todas las empresas o industrias serán afectadas de la misma manera por eventos específicos. Por ejemplo, problemas internos en una empresa, como la mala gestión o el lanzamiento fallido de un producto, pueden afectar negativamente a las acciones de esa empresa, pero estos problemas no necesariamente impactarán a otras empresas o industrias.

En resumen, mientras que el riesgo sistemático afecta a todo el mercado y no puede ser mitigado a través de la diversificación, el riesgo no sistemático es específico de empresas o industrias individuales y puede ser reducido mediante una estrategia de diversificación adecuada.

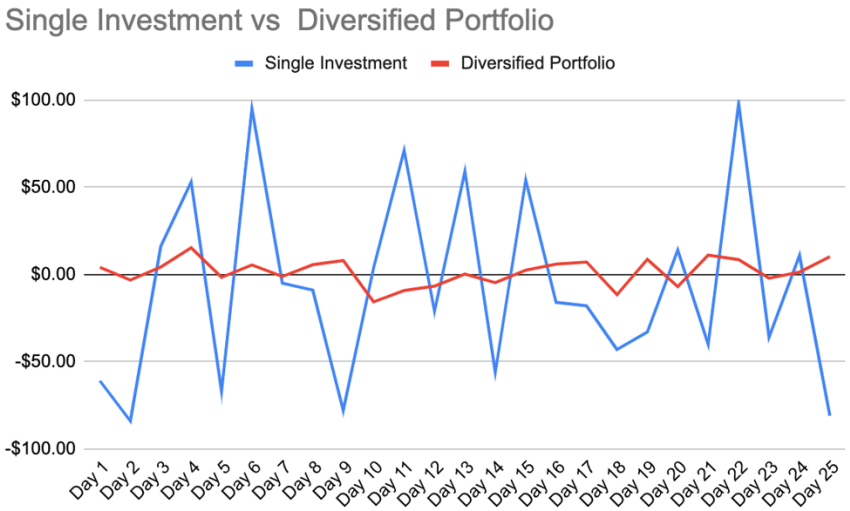

Figura 2.2: Inversión Única vs Diversificada

Pero, aquí surge un debate. La concentración, según algunos expertos en inversiones, podría ser más fructífera que la diversificación excesiva. Argumentan que un portafolio de unas pocas empresas seleccionadas cuidadosamente puede superar a uno más diversificado. La clave está en la selección meticulosa y el conocimiento profundo de cada inversión realizada.

Históricamente, el S&P 500 ha demostrado que las rentabilidades de largo plazo provienen de una minoría de acciones que exhiben rendimientos extraordinarios, mientras que la mayoría solo añade un peso marginal.

Entonces, ¿por qué incrementar la carga de una cartera con activos subóptimos cuando es posible seleccionar exclusivamente aquellos de alto rendimiento?

En la práctica, esto se traduce en elegir menos empresas, pero más prometedoras. Un portafolio concentrado, compuesto por empresas con fundamentos sólidos y potencial de crecimiento, podría, en teoría, reducir la volatilidad y aumentar las ganancias en comparación con uno más diversificado, pero con una selección menos acertada.

La diversificación, parece tener un límite en su eficacia, y más allá de cierto punto, puede diluir las ganancias potenciales más que proteger contra las pérdidas.

El gráfico 2.3 representa la variabilidad en un portafolio compuesto por diversas cantidades de empresas en comparación con el índice S&P 500. La línea punteada refleja la volatilidad del S&P 500, mientras que la línea continua muestra la volatilidad del portafo-

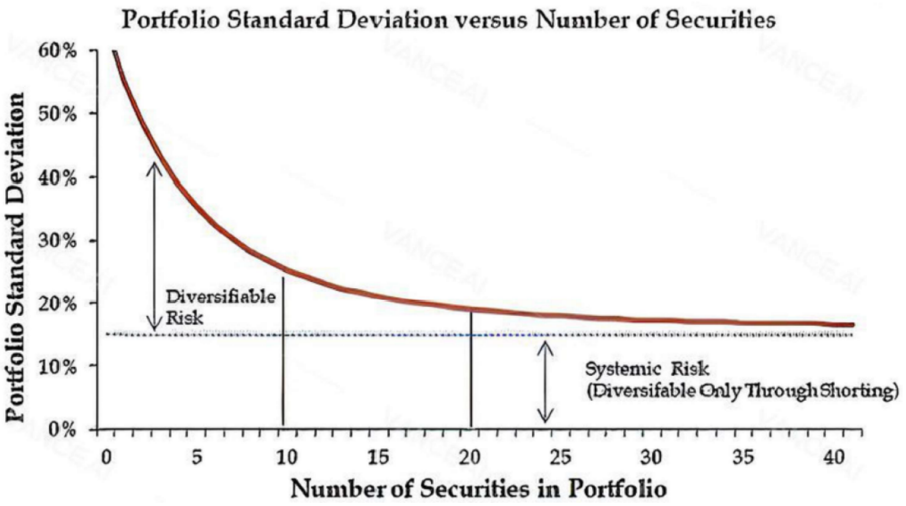

Figura 2.3: Desviación Típica de las Carteras en función del número de empresas en el portafolio

lio, variando según el número de empresas incluidas, que se presenta en el eje horizontal. En el eje vertical se representa la medida de esta volatilidad.

Este enfoque [19] desafía la noción tradicional y sugiere que la verdadera habilidad en la inversión no reside en esparcir el riesgo indiscriminadamente, sino en concentrar el capital en oportunidades verdaderamente valiosas. La diversificación protege del error, pero la concentración se enfoca en el acierto. Así, un inversor informado y selectivo puede, con menos, lograr más.

La discusión sobre la eficacia de la diversificación frente a la concentración en una cartera de inversión es central en la teoría financiera moderna [6]. Mientras la diversificación es tradicionalmente ensalzada por su capacidad para mitigar riesgos, algunos inversores de valor plantean una perspectiva alternativa, afirmando que una cartera altamente concentrada puede, de hecho, ofrecer mejores rendimientos ajustados por riesgo.

La idea central detrás de esta perspectiva es que la diversificación extrema puede llevar a una mediocridad en los rendimientos. Incluso Warren Buffett ha criticado la diversificación excesiva [20], sugiriendo que "la diversificación es una protección contra la ignorancia", implicando que un conocimiento profundo y un enfoque selectivo pueden ser más provechosos.

Además, estudios históricos han demostrado que las carteras altamente concentradas, que poseen una selección cuidadosa de acciones, no solo alcanzan, sino que a menudo superan, los rendimientos del índice S&P 500 [27]. Esto contradice la noción convencional de que aumentar el número de inversiones en una cartera reduce el riesgo y la volatilidad.

Por ejemplo, un estudio de 1970 realizado por Lawrence Fisher y James H. Lorie [12] encontró que una cartera aleatoriamente compuesta de 30 empresas exhibía una volatilidad y desviación estándar cercanas a las del índice S&P 500. Este resultado sugiere que, más allá de un cierto número de acciones, los beneficios de la diversificación en términos de reducción de la volatilidad disminuyen.

El riesgo diversificable se puede mitigar añadiendo diferentes acciones al portafolio, pero el riesgo sistemático, inherente al mercado, persiste independientemente de cuántas acciones contenga el portafolio. La clave, entonces, no está en la cantidad, sino en la calidad y la interrelación de las inversiones seleccionadas.

Con estos conceptos en mente, un inversor tiende a construir un portafolio más concentrado, enfocándose en empresas que no solo tienen un sólido rendimiento histórico

	1 empresa	15 empresas	30 empresas	60 empresas	SP500
Desviación Típica	45%	16.50%	15.40%	15.20%	14.50%
Error de Rastreo	45%	8.1%	6.2%	5.3%	0%

Figura 2.4: Simulación Desviación Típica y Error de Rastreo en función del número de empresas

sino que también muestran potencial de crecimiento y estabilidad financiera. El objetivo es seleccionar empresas que, en conjunto, ofrecen un balance optimizado entre rendimiento y riesgo, sin la necesidad de una diversificación excesiva que pueda diluir los potenciales altos rendimientos.

Este enfoque concentrado destaca la importancia de la selección y el análisis riguroso, defendiendo que un número menor de inversiones cuidadosamente escogidas puede resultar en un desempeño superior, refutando la idea de que más siempre es mejor en el contexto de la diversificación de una cartera de inversión.

2.4 Justificación de la Selección de Modelos Predictivos

En la elaboración de este Trabajo Fin de Grado, se seleccionaron varios modelos predictivos basados en sus características únicas y su aplicabilidad a los datos financieros [28]. A continuación, se justifica la elección de cada modelo y cómo su implementación puede contribuir a alcanzar los objetivos de este estudio.

2.4.1. Regresión Lineal y Ridge

La Regresión Lineal [8] es fundamental para establecer relaciones lineales entre variables y es ampliamente utilizada por su simplicidad y eficacia en la predicción de valores continuos. La elección de este modelo se justifica por su capacidad para proporcionar un punto de partida sólido y comprensible para las predicciones financieras. Por otro lado, el modelo Ridge extiende la regresión lineal al introducir una penalización L2 que controla la complejidad del modelo, evitando el sobreajuste a los datos de entrenamiento. Este modelo es especialmente útil en situaciones donde hay una alta correlación entre las variables predictoras, una condición común en los datos financieros.

2.4.2. Lasso y Regresión ARD

El modelo Lasso [14] implementa una penalización L1 que tiene la ventaja adicional de realizar la selección de variables automáticamente al reducir los coeficientes de las variables menos importantes a cero. Esta propiedad es invaluable cuando se manejan datasets con un gran número de variables, permitiendo simplificar el modelo y mejorar la interpretación de los resultados. La Regresión ARD (Automatic Relevance Determination), similar a Lasso pero desde un enfoque bayesiano, permite ajustar la complejidad del modelo y proporciona una medida de incertidumbre en las predicciones, lo cual es crucial para la toma de decisiones en el ámbito financiero.

2.4.3. Random Forest y XGBoost

Random Forest [7] es un modelo de ensamble que utiliza múltiples árboles de decisión para obtener una predicción más estable y menos susceptible a la variabilidad de los

datos, lo que lo hace robusto ante el sobreajuste. XGBoost mejora el algoritmo de boosting al incorporar regularización, lo que mejora considerablemente su capacidad para generalizar sobre datos no vistos. Estos modelos son elegidos por su eficacia demostrada en capturar relaciones no lineales y su capacidad para manejar datos de alta dimensionalidad y complejidad, características típicas del mercado financiero.

2.4.4. Redes Neuronales

Las Redes Neuronales Multicapa [18] son capaces de modelar interacciones complejas y no lineales entre variables. Su selección se justifica por la flexibilidad que ofrecen en la modelización de patrones complejos y su éxito en numerosas aplicaciones de series temporales financieras. Además, la capacidad de las redes neuronales para aprender representaciones a diferentes niveles de abstracción permite explorar profundamente los datos en busca de patrones que otros modelos lineales podrían pasar por alto.

Estos modelos fueron elegidos no solo por su robustez y precisión sino también por su adaptabilidad a diferentes tipos de estructuras de datos , lo que es esencial para responder a los desafíos presentados por los mercados financieros dinámicos y a menudo impredecibles. La combinación de estos modelos permite construir un enfoque comprensivo y poderoso para la predicción de rendimientos financieros, maximizando la posibilidad de superar al índice S&P 500 como se plantea en los objetivos de este estudio.

CAPÍTULO 3
Metodología

3.1 Recopilación de Datos Financieros

3.1.1. Selección de Fuentes de Datos

Las fuentes de datos para la recolección de información financiera fueron seleccionadas por su fiabilidad y actualización continua. Se utilizó, principalmente, la plataforma StockAnalysis[1], conocida por proporcionar datos financieros detallados de empresas listadas en bolsa. Este tipo de plataformas son esenciales para obtener datos precisos y actuales que son cruciales para el análisis financiero.

Se recopilaron 2 bases de datos para construir 2 modelos diferentes.

La primera base de datos consiste solamente en datos anualizados. Desde el año 2000 al 2023. El objetivo con estos datos es construir un modelo con un horizonte temporal en las predicciones de 5 años.

La segunda base de datos compuesta con datos trimestrales desde 2010 a 2023. El objetivo es construir un modelo con un horizonte temporal mucho más reducido. De 1 año.

Sin embargo, para ambas bases de datos el proceso de recolección fue idéntico. Se detalla a continuación.

3.1.2. Proceso de Recolección de Datos

La recolección de datos se llevó a cabo mediante scripts automatizados en Python[2], lo que incluyó la autenticación segura en las plataformas de datos y la extracción sistemática de información financiera relevante.

[1] https://stockanalysis.com/
[2] https://www.python.org/

```python

import requests
from bs4 import BeautifulSoup

# URL de la página de Wikipedia
url = 'https://es.wikipedia.org/wiki/Anexo:Compañías_del_S%26P_500'
response = requests.get(url)

# Verificar si la solicitud fue exitosa
if response.status_code == 200:
    soup = BeautifulSoup(response.content, 'html.parser')
    table = soup.find('table', {'class': 'wikitable'})
    rows = table.find_all('tr')
    tickers_before_2016 = []

    for row in rows[1:]:
        cols = row.find_all('td')
        if len(cols) > 6:
            ticker = cols[0].text.strip()
            date_incorporated = cols[6].text.strip()
            if date_incorporated < '2018-01-01':
                tickers_before_2016.append(ticker)
else:
    print("Error al obtener la página")
```

Listing 3.1: Proceso de Recolección de Datos

3.1.3. Autenticación y Acceso

Utilizamos métodos de autenticación seguros que incluyeron la validación de credenciales y el uso de tokens de acceso. Esto aseguró que solo los usuarios autorizados pudieran acceder a las bases de datos y que toda la información sensible se mantuviera protegida durante el proceso de recopilación y análisis de datos.

```python
import requests

# Credenciales de acceso
login_url = 'https://www.stockanalysis.com/login/'
payload = {'username': 'user@example.com', 'password': 'securepassword'}

# Iniciar sesión para autenticación
session = requests.Session()
session.post(login_url, data=payload)
```

Listing 3.2: Autenticación

3.1.4. Extracción Automatizada de Datos

Se utilizaron scripts automatizados en Python para extraer datos de manera sistemática. Se emplearon bibliotecas como `requests`[3] y `BeautifulSoup`[4] para automatizar la extracción de información de las páginas web de cada empresa, asegurando así la eficiencia y la repetibilidad del proceso.

```python
from bs4 import BeautifulSoup

# Funcion para extraer datos de una empresa
```

[3] https://pypi.org/project/requests/
[4] https://pypi.org/project/beautifulsoup4/

```
def extract_data(url):
    page = session.get(url)
    soup = BeautifulSoup(page.content, 'html.parser')
    return soup

# Uso de la función con una URL específica
data = extract_data('https://www.stockanalysis.com/stocks/AAPL/financials/')
```

Listing 3.3: Función webscraping

3.1.5. Limpieza y Preparación de Datos

Una vez recogidos los datos, fueron limpiados para eliminar cualquier inconsistencia o valores faltantes. Este paso es crucial para asegurar la precisión de los modelos predictivos. La limpieza incluyó la normalización de formatos y la consolidación de los datos en un conjunto único para su análisis.

```
import pandas as pd

# Limpieza de datos en un DataFrame
def clean_data(df):
    df.fillna(method='ffill', inplace=True)  # Llenar valores faltantes
    df.replace({'%': ''}, regex=True, inplace=True)  # Eliminar símbolos no
        deseados
    return df

# Aplicar la limpieza a un DataFrame de ejemplo
df = pd.DataFrame(data)
clean_df = clean_data(df)
```

Listing 3.4: Limpieza de datos

3.1.6. Almacenamiento de Datos

Para que el acceso y el uso de los datos durante la fase de modelado fueran más fáciles, los almacenamos de manera organizada. Usamos bases de datos y sistemas de archivos que funcionan bien con herramientas de análisis avanzadas. Esto nos permitió trabajar de manera eficiente durante todo el proceso de análisis. Mantener los datos organizados y accesibles fue clave para optimizar todo el flujo de trabajo analítico.

```
# Guardar DataFrame en un archivo CSV
df.to_csv('financial_data.csv', index=False)
```

3.2 Métodos de Selección de Características

3.2.1. Técnicas de Preprocesamiento

En la fase de preprocesamiento, los datos fueron sometidos a varias técnicas para garantizar su calidad y coherencia. Se llevaron a cabo tareas como la eliminación de valores faltantes y la corrección de datos atípicos [21]. Además, se normalizaron las variables para asegurar que estuvieran en un formato consistente, lo que facilitó el análisis y mejoró la precisión de los modelos predictivos.

```
import pandas as pd

```

```
# Limpieza de datos en un DataFrame
def clean_data(df):
    df.fillna(method='ffill', inplace=True)  # Llenar valores faltantes
    df.replace({'%': ''}, regex=True, inplace=True)  # Eliminar símbolos no deseados
    return df

# Aplicar la limpieza a un DataFrame de ejemplo
df = pd.DataFrame(data)
clean_df = clean_data(df)
```

Listing 3.5: Limpieza de datos en un DataFrame

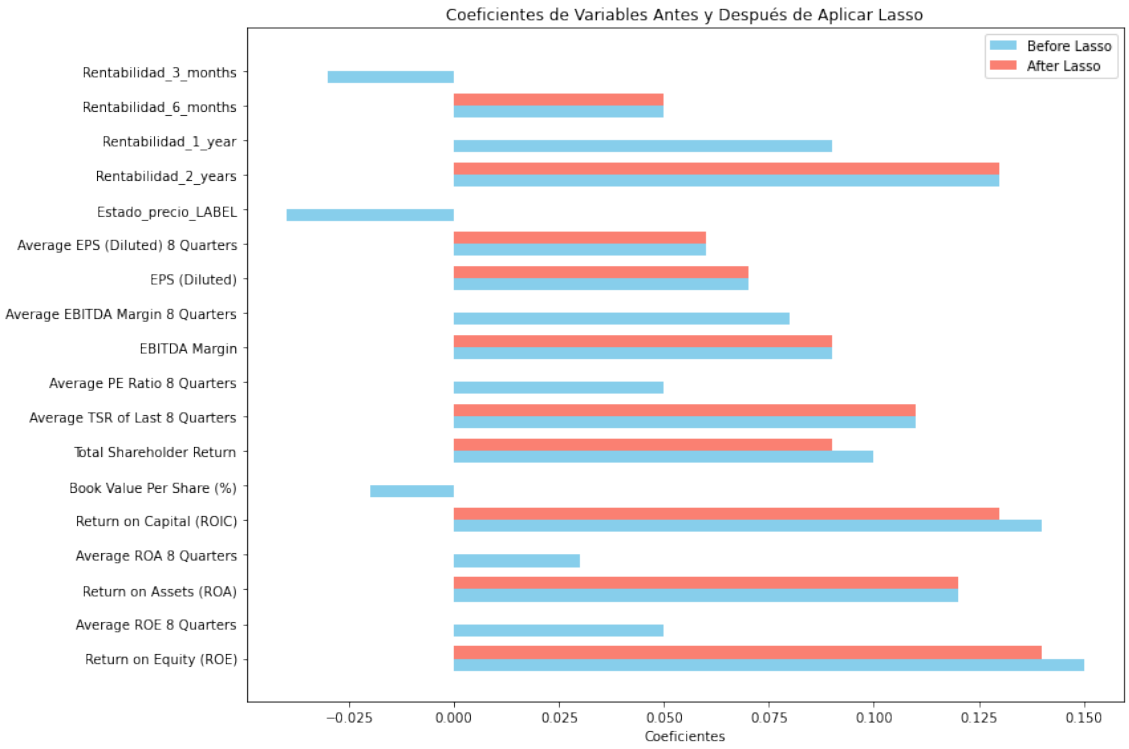

Figura 3.1: Aplicación de Lasso

3.2.2. Selección de Características Relevantes

Para seleccionar las características más relevantes en nuestro estudio, empleamos una serie de técnicas avanzadas que nos permitieron identificar las variables clave que contribuyen a la precisión de los modelos predictivos. Este proceso incluyó varias etapas, comenzando con un preprocesamiento riguroso de los datos financieros recopilados. Este paso fue fundamental para asegurar la limpieza y normalización de los datos, eliminando valores atípicos y completando datos faltantes mediante métodos estadísticos.

La Figura 3.2 muestra la matriz de correlaciones entre diversas métricas financieras. Se observa una alta correlación entre el Return on Assets (ROA) y el Return on Equity (ROE), lo cual indica que ambos podrían estar capturando información similar sobre la rentabilidad de la empresa. Por otro lado, métricas como el Book Value Per Share (%) presentan una baja correlación con la mayoría de las otras variables, sugiriendo que podrían añadir información complementaria al modelo. Este análisis es crucial para identificar las relaciones entre variables y determinar cuáles pueden ser redundantes o añadir valor al modelo predictivo. La comprensión de estas correlaciones ayuda a mejorar la selección de características y la precisión del modelo.

Por otro lado, en la Figura 3.1, se observa cómo la regresión Lasso ajusta los coeficientes de las variables, eliminando aquellas menos significativas y simplificando el modelo. Esto mejora la interpretabilidad y permite centrarse en las variables que realmente aportan valor al rendimiento predictivo del modelo.

Por otro lado, la regresión ARD adoptó un enfoque bayesiano, ajustando la complejidad del modelo y proporcionando una medida de incertidumbre en las predicciones. Esta técnica resultó crucial en el contexto financiero, donde la precisión y la capacidad de interpretar las variables relevantes son esenciales para la toma de decisiones informadas.

Finalmente, llevamos a cabo la optimización de los modelos mediante técnicas de validación cruzada y ajuste de hiperparámetros. La validación cruzada fue esencial para

	Return on Equity (ROE)	Average ROE 8 Quarters	Return on Assets (ROA)	Average ROA 8 Quarters	Return on Capital (ROIC)	Book Value Per Share (%)	Total Shareholder Return	Average TSR of Last 8 Quarters	Average PE Ratio 8 Quarters
Return on Equity (ROE)	1.000000	-0.083967	0.015118	0.003506	-0.019643	-0.002603	0.001531	-0.014076	-0.000753
Average ROE 8 Quarters	-0.083967	1.000000	0.051630	0.122329	0.026624	-0.046387	0.028441	0.059204	0.016799
Return on Assets (ROA)	0.015118	0.051630	1.000000	0.477630	0.753376	-0.400512	0.160564	0.111280	0.049654
Average ROA 8 Quarters	0.003506	0.122329	0.477630	1.000000	0.361635	-0.230729	0.075951	0.281162	0.223196
Return on Capital (ROIC)	-0.019643	0.026624	0.753376	0.361635	1.000000	-0.370360	0.156501	0.119780	0.040915
Book Value Per Share (%)	-0.002603	-0.046387	-0.400512	-0.230729	-0.370360	1.000000	-0.003990	-0.018384	-0.040929
Total Shareholder Return	0.001531	0.028441	0.160564	0.075951	0.156501	-0.003990	1.000000	0.299335	-0.017073
Average TSR of Last 8 Quarters	-0.014076	0.059204	0.111280	0.281162	0.119780	-0.018384	0.299335	1.000000	-0.020588
Average PE Ratio 8 Quarters	-0.000753	0.016799	0.049654	0.223196	0.040915	-0.040929	-0.017073	-0.020588	1.000000
EBITDA Margin	0.009112	0.005144	0.282524	0.096537	0.145958	0.048114	-0.006105	-0.048309	0.059802
Average EBITDA Margin 8 Quarters	0.000608	0.062856	0.062226	0.526658	0.018273	0.047173	-0.009332	0.120826	0.254244
EPS (Diluted)	-0.008136	-0.005088	0.415563	0.206513	0.410578	-0.081361	0.115664	0.073423	0.037460
Average EPS (Diluted) 8 Quarters	-0.016155	0.060414	0.169914	0.571454	0.188402	-0.026978	0.095101	0.259256	0.182077
Estado_precio	-0.019945	0.058126	0.005787	0.490438	0.020008	-0.034010	-0.042827	0.181889	0.226669
LABEL	-0.007508	0.023691	0.010570	0.106923	0.045029	-0.051671	0.023470	0.076110	0.009948
Rentabilidad_2_years	-0.027330	0.010225	0.108404	0.251254	0.085582	-0.143545	-0.097829	-0.001530	0.128162
Rentabilidad_1_year	-0.033803	-0.004753	0.059217	0.075701	0.056218	-0.153740	-0.097064	0.015205	0.058809
Rentabilidad_6_months	-0.019913	-0.006929	0.035869	0.063214	0.034848	-0.133051	-0.065546	0.030490	0.041660
Rentabilidad_3_months	-0.028070	0.008166	0.015670	0.060796	0.014014	-0.098840	-0.040802	0.038942	0.031868

Figura 3.2: Análisis de correlaciones

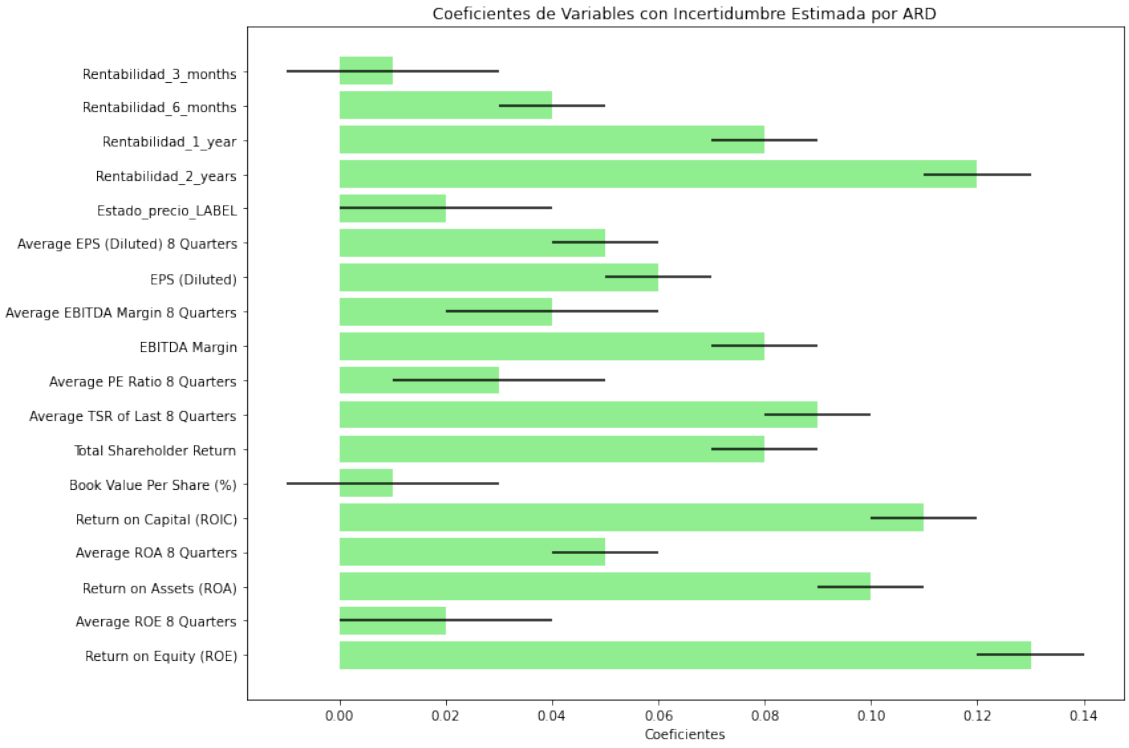

Figura 3.3: Coeficientes de Variables con Incertidumbre Estimada por ARD

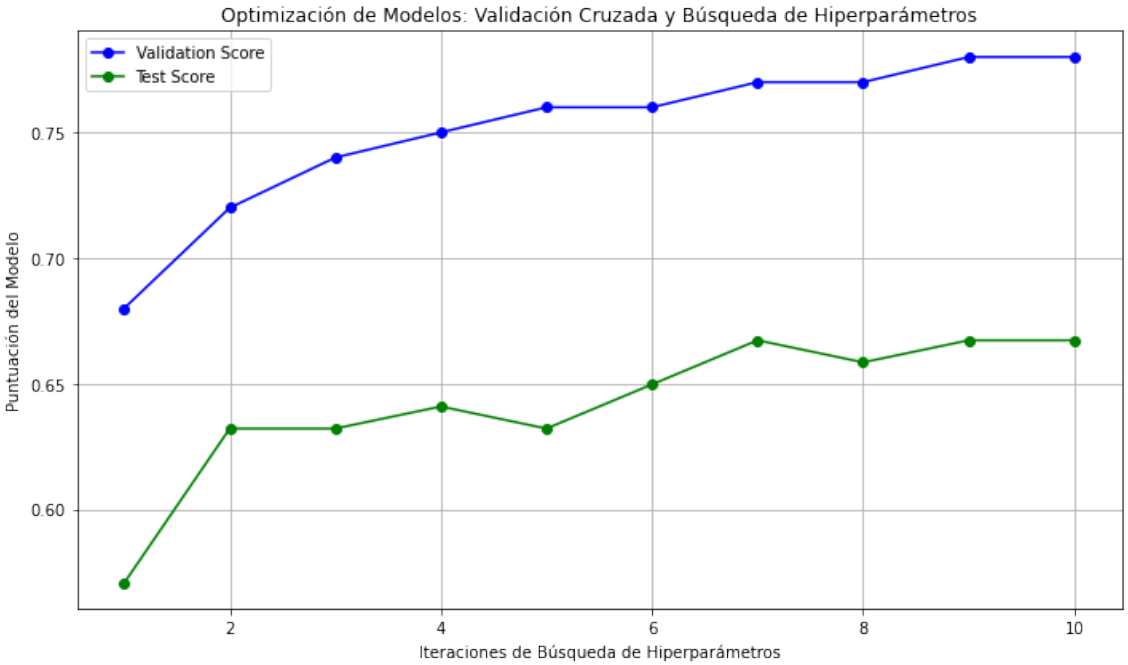

Figura 3.4: Validacion cruzada

evaluar la capacidad de generalización de los modelos, ayudando a prevenir el sobreajuste. Utilizamos varios esquemas de validación, como k-fold y leave-one-out, para obtener una estimación precisa del rendimiento del modelo en datos no vistos. En la Figura 3.4, se observa cómo los puntajes de validación y prueba del modelo mejoran con el aumento de iteraciones de búsqueda de hiperparámetros. La puntuación de validación muestra una mejora continua y significativa, estabilizándose después de aproximadamente ocho iteraciones, lo que indica que el modelo se vuelve más robusto y ajustado. Por otro lado, la puntuación de prueba también mejora, aunque de manera más moderada, sugiriendo que la validación cruzada y la optimización de hiperparámetros son esenciales para garantizar la generalización del modelo a nuevos datos.

3.3 Construcción y Optimización de Modelos Predictivos

3.3.1. Implementación de Modelos de Aprendizaje Automático

Para la construcción y optimización de nuestros modelos predictivos, implementamos y evaluamos diversos modelos de aprendizaje automático. Entre ellos se incluyeron la regresión lineal, la regresión Ridge, la regresión Lasso, Random Forest, XGBoost y redes neuronales. La elección de estos modelos se basó en su capacidad para manejar datos financieros que son inherentemente de alta dimensionalidad y complejidad.

Cada uno de estos modelos fue seleccionado por sus características únicas y su aplicabilidad al análisis de datos financieros. La regresión lineal y Ridge proporcionaron un punto de partida sólido, ofreciendo simplicidad y facilidad de interpretación. La regresión Lasso y ARD, por su parte, fueron cruciales para la selección de características y el manejo de datasets con muchas variables.

Los modelos de ensamble como Random Forest y XGBoost [4] demostraron ser particularmente eficaces para capturar relaciones complejas y no lineales entre las variables, lo que es típico en los mercados financieros. Finalmente, las redes neuronales multicapa

ofrecieron la flexibilidad necesaria para modelar interacciones complejas y no lineales, lo que resultó en una mejora significativa de las predicciones.

```
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.ensemble import RandomForestRegressor
import xgboost as xgb
from sklearn.neural_network import MLPRegressor

# Ejemplo de implementación de modelos
models = {
    'Linear Regression': LinearRegression(),
    'Ridge Regression': Ridge(),
    'Lasso Regression': Lasso(),
    'Random Forest': RandomForestRegressor(),
    'XGBoost': xgb.XGBRegressor(),
    'Neural Network': MLPRegressor()
}

for name, model in models.items():
    model.fit(X_train, y_train)
    predictions = model.predict(X_test)
    print(f"{name} - Mean Squared Error: {mean_squared_error(y_test,
        predictions)}")
```

Listing 3.6: Implementación de modelos

3.3.2. Optimización de Modelos

Para ajustar y mejorar el rendimiento de los modelos, se aplicaron diversas técnicas de optimización. Entre ellas, se utilizó la validación cruzada y la búsqueda de hiperparámetros, lo que permitió afinar los modelos y asegurar su robustez. Estas técnicas fueron fundamentales para garantizar que los modelos puedan generalizar correctamente cuando se enfrentan a nuevos datos.

3.4 Evaluación de Modelos

3.4.1. Métricas de Evaluación

Para medir el rendimiento de los modelos, se utilizaron varias métricas clave. Se consideraron el error cuadrático medio (MSE) [17], el coeficiente de determinación (R^2) [26] y la precisión en la predicción de retornos. Estas métricas ofrecieron una evaluación detallada de la eficacia de cada modelo.

El **Error Cuadrático Medio (MSE)** se define como:

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

donde y_i son los valores reales y $\hat{y}_i$ son los valores predichos por el modelo.

El **Coeficiente de Determinación (R^2)** se calcula como:

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

donde $\bar{y}$ es la media de los valores reales.

Además de estas métricas, se evaluó la precisión en la predicción de retornos financieros. Para ello, se compararon los valores predichos de retorno contra los valores reales. El retorno (R) se calcula como:

$$R = \frac{P_{final} - P_{inicial}}{P_{inicial}}$$

donde P_{final} es el precio final del activo y $P_{inicial}$ es el precio inicial. Esta métrica es fundamental para la optimización de portafolios desde un punto de vista práctico, ya que permite evaluar directamente la eficacia del modelo en términos de ganancias económicas. La comparación de valores predichos y reales es crucial para entender la precisión y aplicabilidad de los modelos en situaciones de mercado real.

```
from sklearn.metrics import mean_squared_error, r2_score

# Evaluación de modelos
for name, model in models.items():
    predictions = model.predict(X_test)
    mse = mean_squared_error(y_test, predictions)
    r2 = r2_score(y_test, predictions)
    print(f"{name} - MSE: {mse}, R^2: {r2}")
```

Listing 3.7: Evaluación de modelos

3.4.2. Validación Cruzada

La validación cruzada [24] se utilizó para evaluar la capacidad de generalización de los modelos. Este método implica dividir los datos en múltiples subconjuntos y entrenar los modelos en diferentes combinaciones de estos subconjuntos para obtener una estimación más precisa de su rendimiento.

```
from sklearn.model_selection import cross_val_score

# Validación cruzada
for name, model in models.items():
    scores = cross_val_score(model, X, y, cv=5)
    print(f"{name} - Cross-Validation Score: {scores.mean()}")
```

Listing 3.8: Validación cruzada

CAPÍTULO 4

Implementación de Estrategias de Inversión basadas en Ciencia de Datos

4.1 Descripción de las Estrategias Desarrolladas

Ahora, detallaremos las estrategias de inversión desarrolladas utilizando modelos predictivos avanzados. Estas estrategias están diseñadas para aprovechar las capacidades del aprendizaje automático y técnicas de análisis cuantitativo, con el objetivo de optimizar el rendimiento de las carteras de inversión. A continuación, se presentan las estrategias implementadas, cada una con su enfoque específico y el código correspondiente.

Estas estrategias se implementaran de forma idéntica en ambas bases de datos (anualizadas y trimestrales) para construir 2 modelos.

4.1.1. Estrategia Basada en Regresión Lineal y Ridge

La primera estrategia emplea modelos de regresión lineal y regresión Ridge, conocidos por su simplicidad y capacidad de interpretación. Estos modelos ayudan a identificar relaciones lineales entre diversas características financieras y los rendimientos futuros de las acciones.

- **Selección de Características**: Se identificaron las variables financieras clave, como ROE, ROA, y el ratio PE, que son cruciales para predecir los rendimientos.
- **Entrenamiento del Modelo**: Los modelos fueron entrenados con datos históricos, ajustando parámetros para minimizar el error cuadrático medio (MSE).
- **Predicción y Selección de Activos**: Los modelos predecían los rendimientos futuros, y se seleccionaron las acciones con las mejores perspectivas ajustadas por riesgo para incluirlas en la cartera.

```
from sklearn.linear_model import LinearRegression, Ridge

# Entrenamiento de los modelos
linear_model = LinearRegression()
ridge_model = Ridge(alpha=1.0)
linear_model.fit(X_train, y_train)
ridge_model.fit(X_train, y_train)
```

```
8
9  # Predicción y selección de acciones
10 linear_predictions = linear_model.predict(X_test)
11 ridge_predictions = ridge_model.predict(X_test)
12 selected_stocks_linear = X_test[linear_predictions > linear_predictions.mean()]
13 selected_stocks_ridge = X_test[ridge_predictions > ridge_predictions.mean()]
```

4.1.2. Estrategia Basada en Lasso y Regresión ARD

La segunda estrategia se basa en los modelos Lasso y ARD, que son excelentes para la selección automática de características y el manejo de datasets con alta dimensionalidad.

- **Selección Automática de Características**: El modelo Lasso reduce las variables menos importantes a cero, simplificando así el modelo.
- **Entrenamiento del Modelo**: El modelo ARD, que adopta un enfoque bayesiano, se entrenó con las características seleccionadas para mejorar la precisión en la predicción de rendimientos.
- **Construcción de la Cartera**: Se utilizó el modelo para predecir rendimientos y seleccionar las mejores acciones para la cartera.

```
1  from sklearn.linear_model import Lasso, ARDRegression
2
3  # Entrenamiento y selección de características
4  lasso_model = Lasso(alpha=0.01)
5  lasso_model.fit(X_train, y_train)
6  selected_features = X_train.columns[(lasso_model.coef_ != 0)]
7
8  # Entrenamiento del modelo ARD
9  ard_model = ARDRegression()
10 ard_model.fit(X_train[selected_features], y_train)
11
12 # Predicción y selección de acciones
13 ard_predictions = ard_model.predict(X_test[selected_features])
14 selected_stocks_ard = X_test[ard_predictions > ard_predictions.mean()]
```

Listing 4.1: Entrenamiento modelos

4.1.3. Estrategia Basada en Random Forest y XGBoost

La tercera estrategia aprovecha modelos de ensamble como Random Forest y XGBoost, que son particularmente eficaces para manejar relaciones complejas y no lineales entre variables.

- **Entrenamiento del Modelo**: Se entrenaron los modelos de Random Forest y XGBoost con datos históricos, optimizando hiperparámetros para mejorar la precisión.
- **Predicción de Rendimientos**: Los modelos predijeron los rendimientos futuros de las acciones.
- **Construcción de la Cartera**: Se seleccionaron las acciones con mejores predicciones ajustadas por riesgo para incluirlas en la cartera.

```
from sklearn.ensemble import RandomForestRegressor
import xgboost as xgb

# Entrenamiento de los modelos
rf_model = RandomForestRegressor(n_estimators=100)
xgb_model = xgb.XGBRegressor()
rf_model.fit(X_train, y_train)
xgb_model.fit(X_train, y_train)

# Predicción y selección de acciones
rf_predictions = rf_model.predict(X_test)
xgb_predictions = xgb_model.predict(X_test)
selected_stocks_rf = X_test[rf_predictions > rf_predictions.mean()]
selected_stocks_xgb = X_test[xgb_predictions > xgb_predictions.mean()]
```

Listing 4.2: Entrenamiento modelos con NN

4.1.4. Estrategia Basada en Redes Neuronales

La última estrategia se basa en redes neuronales, conocidas por su capacidad para capturar interacciones complejas y patrones no lineales.

- **Diseño de la Red Neuronal**: Se configuró una red neuronal multicapa adecuada para predecir los rendimientos de las acciones.
- **Entrenamiento de la Red**: La red fue entrenada con datos históricos, ajustando sus pesos para minimizar el error de predicción.
- **Predicción y Selección de Activos**: Se aplicó la red para predecir rendimientos futuros y seleccionar las acciones con las mejores perspectivas para la cartera.

```

from sklearn.neural_network import MLPRegressor

# Configuración y entrenamiento de la red neuronal
nn_model = MLPRegressor(hidden_layer_sizes=(100, 100), max_iter=500)
nn_model.fit(X_train, y_train)

# Predicción y selección de acciones
nn_predictions = nn_model.predict(X_test)
selected_stocks_nn = X_test[nn_predictions > nn_predictions.mean()]
```

Listing 4.3: Entrenamiento Red Neuronal

4.1.5. Ejemplo Práctico: Análisis y Selección de Activos

Para ilustrar cómo se aplican estas estrategias, consideremos un ejemplo práctico de análisis y selección de activos. Supongamos que tenemos un dataset con los siguientes datos financieros de empresas:

```
import pandas as pd
import numpy as np

# Simulación de datos financieros
data = {
    'ticker': ['AAPL', 'MSFT', 'GOOGL', 'AMZN', 'FB'],
    'ROE': [0.30, 0.25, 0.20, 0.15, 0.10],
```

```
    'ROA': [0.15, 0.12, 0.10, 0.08, 0.05],
    'ROIC': [0.25, 0.20, 0.18, 0.15, 0.12],
    'PE_ratio': [25, 30, 28, 22, 20],
    'EPS': [5.0, 4.8, 4.6, 4.4, 4.2]
}

df = pd.DataFrame(data)
```

Listing 4.4: Simulación de datos financieros

Para construir estos modelos se seleccionaron los datos anualizados de 2000 a 2020. Se seleccionó como conjunto de entrenamiento todos los datos del año 2000 al 2019. Y de Prueba los de 2020.

Primero, se prepararon los datos seleccionando solo las columnas numéricas y rellenando los valores faltantes con ceros, como se muestra en el Código 4.5. Luego, se separaron las características (X2) de la etiqueta (y2) y se escalaron los datos utilizando `StandardScaler` de `sklearn`, normalizando así las características para tener una media de 0 y una desviación estándar de 1.

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler

# Selecciona solo columnas numéricas y rellena valores faltantes con 0
df_numeric2 = df2.select_dtypes(include=[np.number]).fillna(0)

# Separa características (X2) y etiqueta (y2)
X2 = df_numeric2.drop('LABEL', axis=1)
y2 = df_numeric2['LABEL']

# Escala los datos
scaler = StandardScaler()
X2 = scaler.fit_transform(X2)
```

Listing 4.5: Preparación y normalización de los datos

A continuación, se usan tres modelos de los anteriores (Se probaron todos los modelos y al final se seleccionarán los que mejor funcionen) para hacer predicciones sobre los datos escalados.

Los resultados de estas predicciones se almacenaron en tres nuevas columnas del DataFrame original (Código 4.6).

```
# Predice usando tres modelos diferentes
y_pred = model.predict(X2)
y_pred2 = model2.predict(X2)
y_pred3 = model3.predict(X2)

# Agrega las predicciones al DataFrame original
df2["PRED"] = y_pred
df2["PRED2"] = y_pred2
df2["PRED3"] = y_pred3
```

Listing 4.6: Predicción utilizando tres modelos

Finalmente, se filtraron las empresas con las mejores predicciones según el tercer modelo (PRED3), seleccionando aquellas que estaban por encima del 85 % superior. De manera similar, se filtraron las empresas según el segundo modelo (PRED2). La intersección de estas dos selecciones se ordenó por las predicciones del primer modelo (PRED) y se seleccionaron las primeras 20 empresas. La media de las etiquetas (LABEL) de las empresas seleccionadas se calculó como se muestra en el Código 4.7.

```
# Filtra las empresas con las mejores predicciones del modelo 3 (por encima del 85% superior)
empresas_seleccionadas = df2[(df2["PRED3"] > df2["PRED3"].quantile(0.85))]
# Filtra las empresas con las mejores predicciones del modelo 2 (por encima del 85% superior)
empresas_seleccionadas2 = df2[(df2["PRED2"] > df2["PRED2"].quantile(0.85))]

# Encuentra la intersección de ambas selecciones y ordena por predicción del modelo 1
empresas_seleccionadas3 = intersection(empresas_seleccionadas,
    empresas_seleccionadas2).sort_values(by='PRED', ascending=False,
    na_position='last')[0:20]

# Calcula la media de las etiquetas (LABEL) de las empresas seleccionadas
media_etiquetas = np.mean(empresas_seleccionadas3["LABEL"])
```

Listing 4.7: Filtrado y selección de las mejores empresas

4.1.6. Resumen

En esta sección, se han presentado diversas estrategias de inversión utilizando modelos predictivos avanzados, como regresión lineal, Ridge, Lasso, ARD, Random Forest, XGBoost y redes neuronales. Estas estrategias no solo explotan las capacidades avanzadas de los modelos de aprendizaje automático, sino que también se adaptan a diferentes escenarios de inversión, optimizando así la construcción de carteras basadas en las predicciones de rendimiento ajustadas por riesgo. Los ejemplos prácticos y el código proporcionado ofrecen una visión clara y práctica de cómo implementar estas estrategias en el análisis y selección de activos financieros.

4.2 Tecnologías y Herramientas Utilizadas

4.2.1. Herramientas de Programación

Para implementar las estrategias de inversión basadas en ciencia de datos, se emplearon diversas herramientas de programación que facilitaron el análisis y modelado de los datos. La principal herramienta utilizada fue Python, un lenguaje de programación altamente versátil y popular en el ámbito de la ciencia de datos y finanzas.

Python no solo es accesible y fácil de aprender, sino que también ofrece una amplia gama de bibliotecas especializadas que permiten realizar desde tareas básicas de manipulación de datos hasta complejas operaciones de aprendizaje automático. A continuación, se describen algunas de las herramientas clave utilizadas en este proyecto:

- **Pandas**[1]: Fundamental para la manipulación y análisis de datos, Pandas permite trabajar con estructuras de datos como *DataFrames*, facilitando la limpieza, transformación y análisis de grandes volúmenes de información financiera.

- **NumPy**[2]: Utilizada para realizar operaciones matemáticas y estadísticas de alto rendimiento, NumPy proporciona soporte para arreglos multidimensionales y una variedad de funciones matemáticas.

[1] https://pandas.pydata.org/
[2] https://numpy.org/

- **Scikit-learn**[3]: Esta biblioteca es una de las más importantes para implementar algoritmos de aprendizaje automático. Scikit-learn ofrece herramientas para la clasificación, regresión, *clustering* y reducción de dimensionalidad, siendo crucial para entrenar y evaluar los modelos predictivos.
- **Matplotlib**[4] **y Seaborn**[5]: Para la visualización de datos, estas bibliotecas permiten crear gráficos y diagramas que ayudan a entender mejor los patrones y relaciones en los datos financieros.

4.2.2. Plataformas y Bibliotecas

Además de las herramientas de programación, se utilizaron diversas plataformas y bibliotecas especializadas que facilitaron el acceso a datos financieros y la implementación de algoritmos avanzados.

- **Jupyter Notebook**: Esta herramienta interactiva permite escribir y ejecutar código de manera fragmentada, ideal para probar y ajustar modelos iterativamente y visualizar resultados inmediatamente. Puedes encontrar más información en su [6].

En conjunto, estas herramientas y plataformas formaron la columna vertebral tecnológica del proyecto, permitiendo desde la recopilación y limpieza de datos hasta la implementación y evaluación de complejos modelos predictivos. La combinación de estas tecnologías no solo facilitó el proceso de desarrollo, sino que también aseguró que las estrategias de inversión fueran robustas y basadas en datos confiables.

[3] https://scikit-learn.org/
[4] https://matplotlib.org/
[5] https://seaborn.pydata.org/
[6] https://jupyter.org

CAPÍTULO 5
Resultados y discusión de estrategias de inversión basadas en Ciencia de Datos

Para ambas Bases de datos los únicos modelos que funcionaban mejor que el índice de referencia fueron:Linear Regression, Lasso y Ridge, como se aprecia en la 5.1.

Fueron estos modelos los empleados finalmente para mostrar los mejores resultados obtenidos siguiendo la metodología presentada en las secciones anteriores.

5.1 Resultados Base de datos anualizada

5.1.1. Comparación de Rendimientos con el S&P 500

En esta sección, se presenta un análisis detallado de los rendimientos a 2 años obtenidos mediante las estrategias de inversión desarrolladas, comparándolos con los del índice S&P 500. Este índice será nuestra referencia o "baseline". Se consideran varios periodos de tiempo para evaluar la consistencia y robustez de las estrategias a lo largo del tiempo.

Se ejecutaron los modelos seleccionados (Lasso, Ridge, Linear Regression) para el año 2020 y nos quedamos con las empresas que estaban en el 15 % de mejores predicciones. Las empresas seleccionadas por los modelos se encuentran en la Figura 5.2.

En la Gráfica 5.3 podemos ver que los resultados muestran un desempeño variado y notablemente superior en comparación con simplemente mantener una posición pasiva en el índice S&P 500.

El análisis revela que, en ciertos periodos, las estrategias basadas en modelos predictivos superan significativamente al S&P 500, lo que sugiere que una selección activa y fundamentada de empresas puede resultar en mayores rendimientos. También podemos ver estos resultados año a año en la Figura 5.4.

La imagen 5.5 muestra las métricas de riesgo y retorno para dos carteras de inversión, Portfolio 1 y "Vanguard 500 Index Investor", durante el período de diciembre de 2021 a diciembre de 2023. Comparando ambas, se observa que "Portfolio 1"supera ampliamente al índice Vanguard en varios aspectos clave, como la media aritmética anualizada (32.84 % frente a 5.58 %), media geométrica anualizada (30.68 % frente a 3.64 %), y ratio de Sharpe (1.36 frente a 0.11). Además, "Portfolio 1"tiene un alpha anualizado significativo de 24.12 %, indicando un rendimiento superior ajustado por riesgo. Sin embargo, Port-

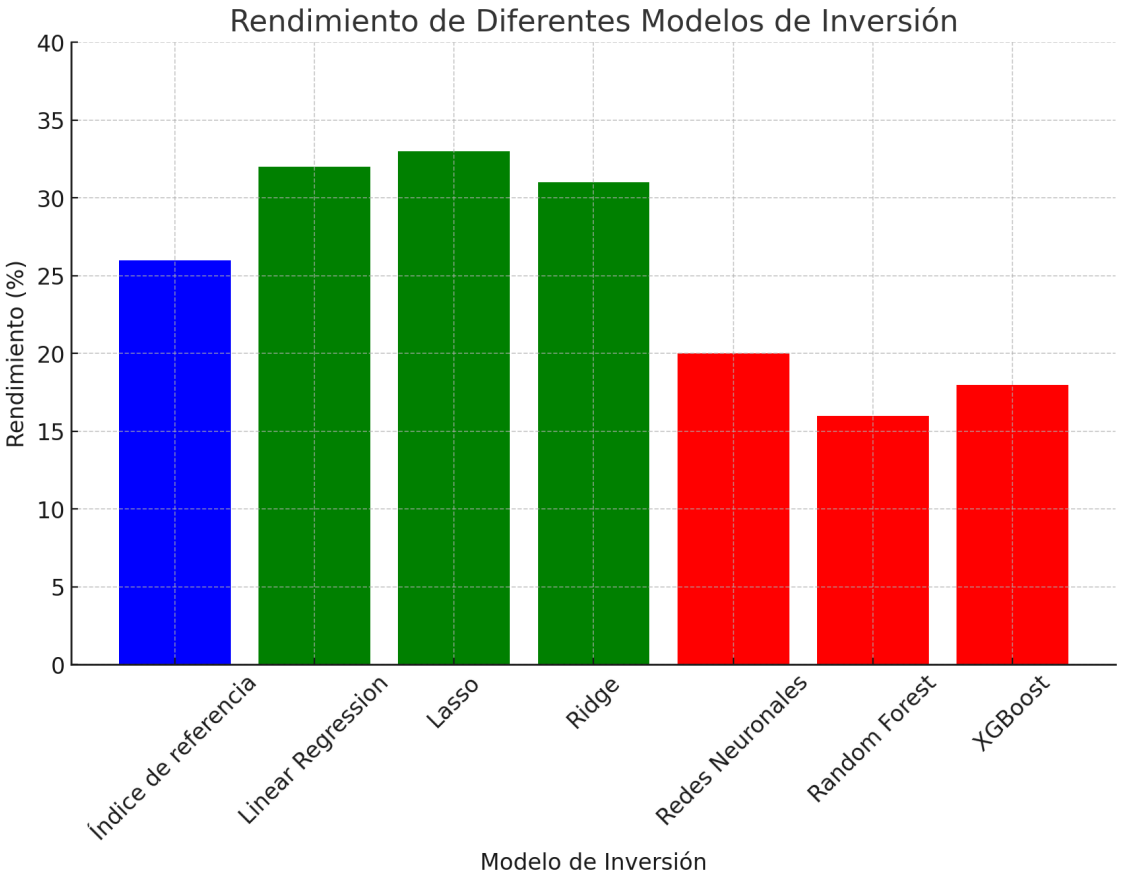

Figura 5.1: Rendimiento difrentes modelos Base de datos Anualizada

Portfolio 1

Ticker	Name	Allocation
VLO	Valero Energy Corporation	5.00%
BHE	Benchmark Electronics, Inc.	5.00%
AXS	Axis Capital Holdings Limited	5.00%
APPF	AppFolio, Inc.	5.00%
UVE	UNIVERSAL INSURANCE HOLDINGS INC	5.00%
CMP	Compass Minerals International, Inc.	5.00%
HY	Hyster-Yale Materials Handling, Inc.	5.00%
MYE	Myers Industries, Inc.	5.00%
DINO	HF Sinclair Corp	5.00%
MLR	Miller Industries, Inc.	5.00%
POWL	Powell Industries, Inc.	5.00%
BELFA	Bel Fuse Inc.	5.00%
SMP	Standard Motor Products, Inc.	5.00%
PECO	Phillips Edison & Co Inc	5.00%
BGSF	BG Staffing Inc	5.00%
GNE	Genie Energy Ltd.	5.00%
PLPC	Preformed Line Products Company	5.00%
RGA	Reinsurance Group of America, Incorporated	5.00%
NCMI	National CineMedia, Inc.	5.00%
GOOG	Alphabet Inc.	5.00%

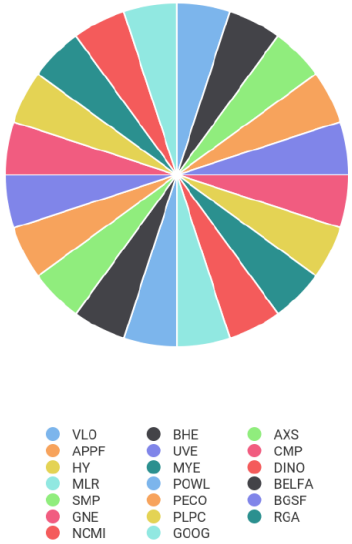

Figura 5.2: Empresas Seleccionadas

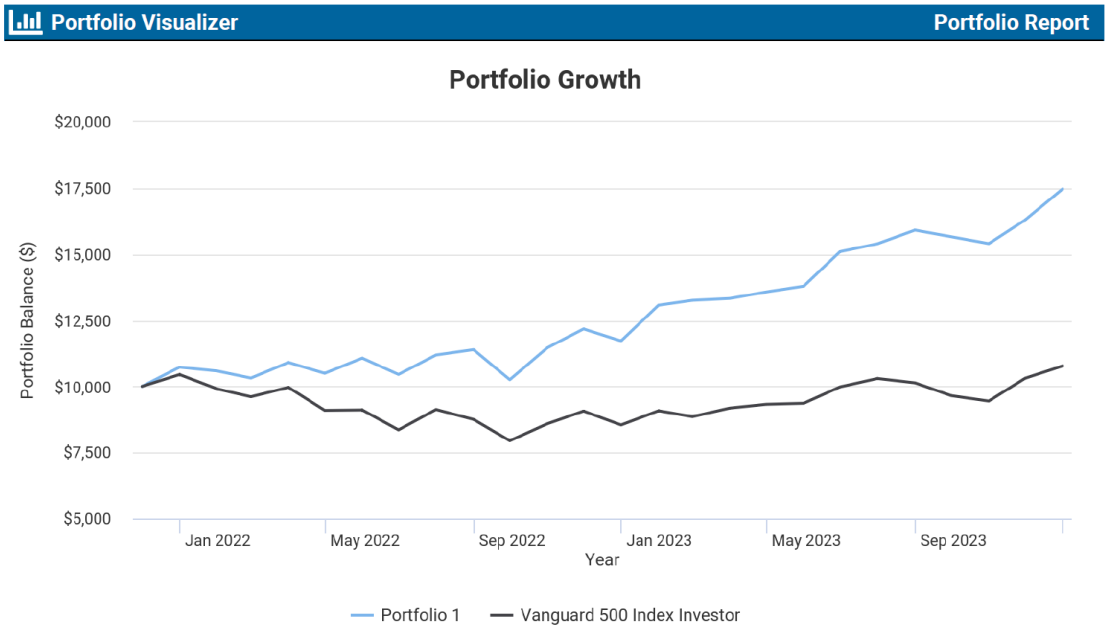

Figura 5.3: Resultados anuales

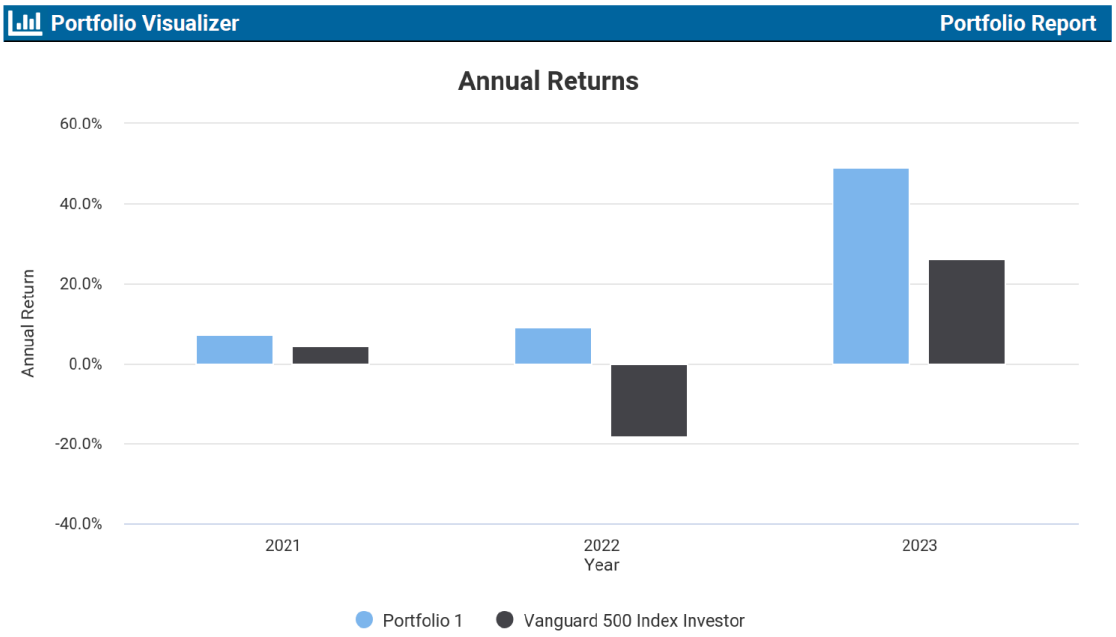

Figura 5.4: Performance año a año

Portfolio Visualizer Portfolio Report

Risk and Return Metrics (Dec 2021 - Dec 2023)

Metric	Portfolio 1	Vanguard 500 Index Investor
Arithmetic Mean (monthly)	2.39%	0.45%
Arithmetic Mean (annualized)	32.84%	5.58%
Geometric Mean (monthly)	2.26%	0.30%
Geometric Mean (annualized)	30.68%	3.64%
Standard Deviation (monthly)	5.43%	5.68%
Standard Deviation (annualized)	18.81%	19.67%
Downside Deviation (monthly)	2.63%	3.75%
Maximum Drawdown	-9.99%	-23.95%
Stock Market Correlation	0.89	1.00
Beta (*)	0.85	1.00
Alpha (annualized)	24.12%	-0.00%
R Squared	78.73%	100.00%
Sharpe Ratio	1.36	0.11
Sortino Ratio	2.70	0.16
Treynor Ratio (%)	29.98	2.15
Active Return	27.04%	N/A
Tracking Error	9.17%	N/A
Information Ratio	2.95	N/A
Skewness	-0.21	-0.13
Excess Kurtosis	-0.26	-1.12
Historical Value-at-Risk (5%)	5.22%	8.64%
Analytical Value-at-Risk (5%)	6.83%	8.89%
Conditional Value-at-Risk (5%)	7.79%	8.98%
Upside Capture Ratio (%)	129.71	100.00
Downside Capture Ratio (%)	46.52	100.00
Positive Periods	17 out of 25 (68.00%)	14 out of 25 (56.00%)
Gain/Loss Ratio	1.40	0.95

(*) Vanguard 500 Index Investor is used as the benchmark for calculations. Value-at-risk metrics are monthly values.

Figura 5.5: Métricas de Riesgo

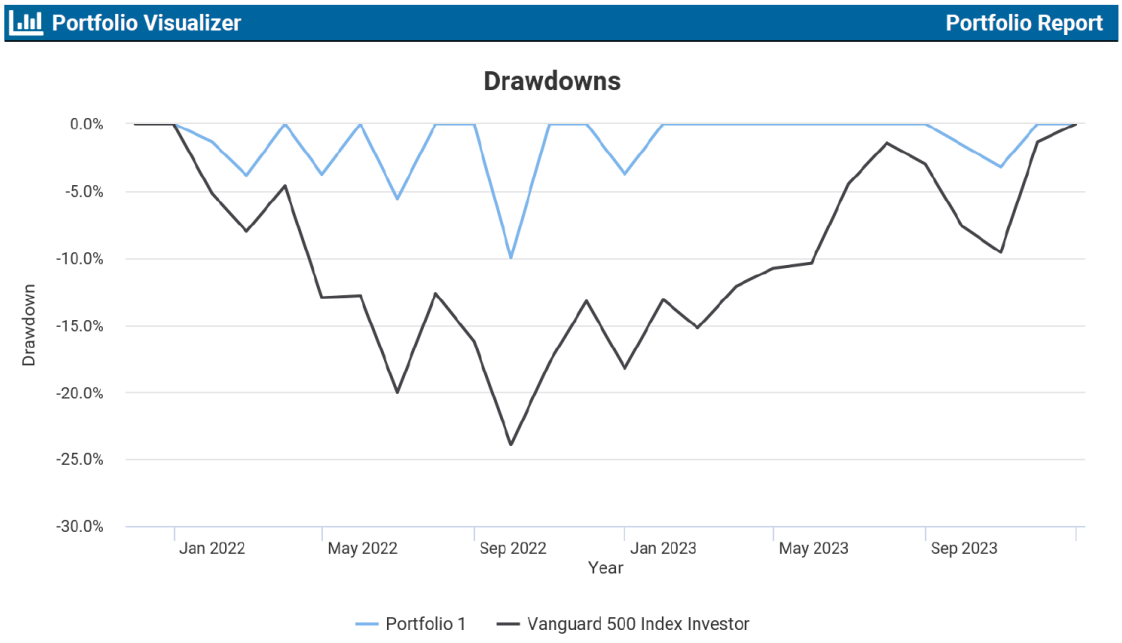

Figura 5.6: Drawdowns

folio 1 también presenta un riesgo considerablemente mayor en términos de desviación estándar anualizada (18.81 % frente a 19.67 %) y un máximo drawdown del -9.99 % en comparación con el -23.95 % del índice. Estas métricas sugieren que, aunque Portfolio 1 ofrece mayores rendimientos potenciales, también implica un riesgo más elevado que el índice de referencia.

Además, la cantidad de empresas seleccionadas se mantiene constante a lo largo del tiempo, lo que permite una comparación directa de los resultados. Esta consistencia en la selección de empresas es crucial para aislar el impacto de las estrategias predictivas en el rendimiento obtenido.

5.1.2. Discusión de Resultados

Aunque los resultados hayan sido muy positivos, debemos ser cautos ya que no tenemos forma de saber si hemos caído en el Sesgo de Supervivencia. Tenemos los datos de empresas que están activas a día de hoy, pero puede darse el caso de que nuestra base de datos no contenga muchas de las empresas que hayan caído.

5.2 Resultados Base de datos trimestrales

5.2.1. Introducción

En esta sección se analizan los resultados obtenidos a partir de la base de datos trimestrales, centrándose en la aplicación de estrategias de inversión dinámicas. Un portafolio dinámico es una estrategia de gestión de inversiones en la que la composición del portafolio se ajusta continuamente en respuesta a los cambios en el mercado y en la economía. A diferencia de los portafolios estáticos, que mantienen una combinación fija de activos a lo largo del tiempo, los portafolios dinámicos buscan aprovechar las oportunidades del mercado y minimizar los riesgos mediante ajustes periódicos en la distribución de los activos.

Un portafolio dinámico se basa en la premisa de que los mercados financieros no son estáticos y que los precios de los activos están influenciados por una variedad de factores que pueden cambiar con el tiempo, como las condiciones macroeconómicas, las políticas gubernamentales, los eventos geopolíticos y las innovaciones tecnológicas. Por lo tanto, un enfoque dinámico permite a los gestores de fondos ajustar sus estrategias para adaptarse a estas condiciones cambiantes, con el objetivo de maximizar el rendimiento ajustado por riesgo.

En la gestión de un portafolio dinámico, se utilizan diferentes modelos predictivos y técnicas cuantitativas para determinar la mejor combinación de activos en cada momento. Estos modelos pueden incluir análisis de series temporales, algoritmos de aprendizaje automático y técnicas de optimización estocástica, entre otros. La idea es identificar patrones y tendencias en los datos históricos para predecir futuros movimientos del mercado y ajustar el portafolio en consecuencia.

En esta parte del trabajo, se implementaron varias estrategias de inversión basadas en modelos predictivos utilizando datos financieros históricos trimestrales. Estas estrategias incluyen el uso de regresiones lineales, modelos de Lasso y ARD, así como técnicas de ensamble como Random Forest y XGBoost. Los resultados de estas estrategias se compararon con un índice de referencia, en este caso, el S&P 500, para evaluar su efectividad. La frecuencia de ajuste trimestral ofrece un balance adecuado entre la capacidad de res-

puesta a los cambios del mercado y la reducción de costos de transacción, que pueden ser significativos si los ajustes se realizan con demasiada frecuencia.

5.2.2. Resultados

Fecha	Baseline	Modelo	Cantidad de Empresas
2018-06-01 00:00:00	7.34	17.34	20
2018-09-01 00:00:00	11.92	36.95	20
2018-12-01 00:00:00	26.74	57.67	20
2019-03-01 00:00:00	32.90	56.59	20
2019-06-01 00:00:00	53.34	68.50	20
2019-09-01 00:00:00	48.76	79.64	20
2019-12-01 00:00:00	33.88	54.39	20
2020-03-01 00:00:00	46.75	60.69	20
2020-06-01 00:00:00	55.95	88.72	20
2020-09-01 00:00:00	37.56	35.11	20
2020-12-01 00:00:00	32.22	13.66	20
2021-03-01 00:00:00	13.99	8.82	20
2021-06-01 00:00:00	-2.20	12.07	20
2021-09-01 00:00:00	3.83	-1.81	20
2021-12-01 00:00:00	13.16	-15.54	20

Tabla 5.1: Comparación de rendimientos entre no hacer nada y el modelo

La tabla 5.1 ofrece una visión integral de cómo las estrategias desarrolladas se desempeñan en diferentes condiciones de mercado, proporcionando una base sólida para futuras investigaciones y ajustes en las metodologías utilizadas. Por tanto, para esta aproximación con datos trimestrales se usaron solo los modelos Lasso, Ridge y Linear Regression.

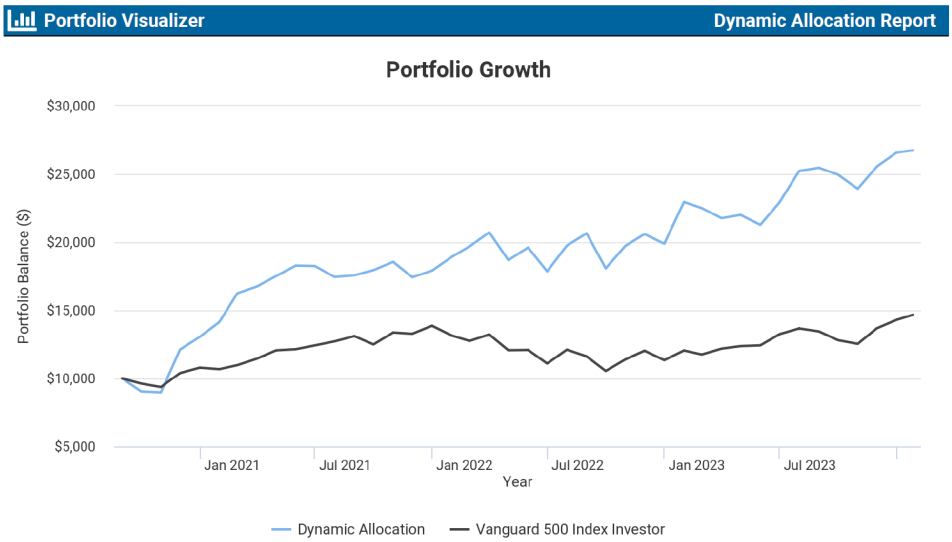

Figura 5.7: Comparación de rendimientos entre el baseline y el modelo en los diferentes periodos de tiempo

En la Figura 5.8 se aprecia como es consistentemente más rentable que el indice de referencia.

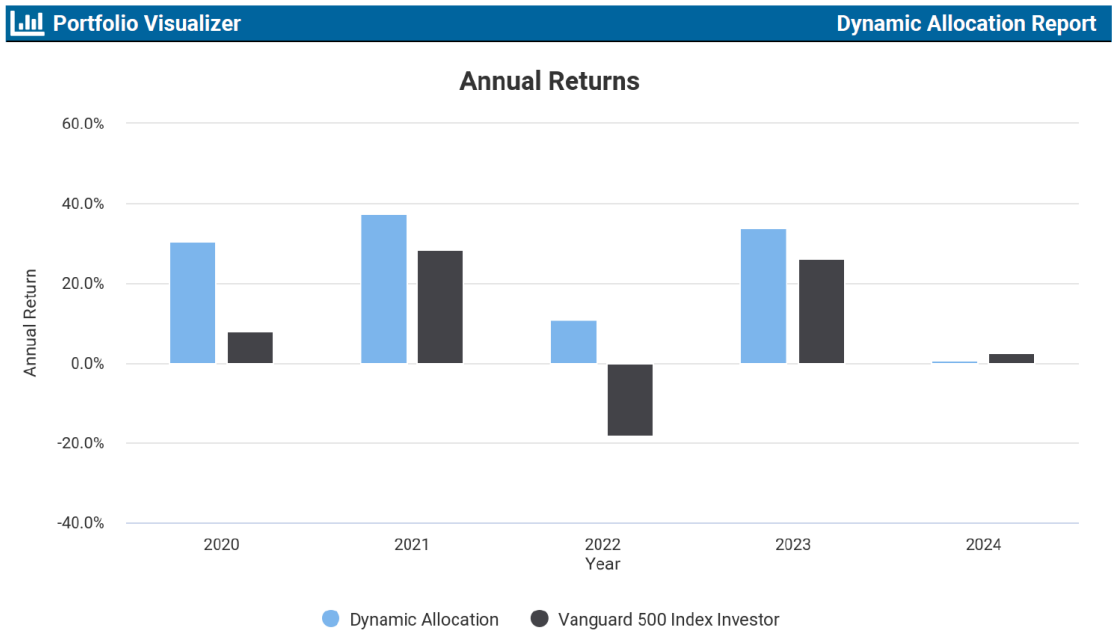

Figura 5.8: Anual Returns de base de datos trimestral

En conclusión, las estrategias de inversión basadas en ciencia de datos y modelos predictivos ofrecen una herramienta poderosa para los inversores, aunque es esencial seguir refinando y adaptando estos modelos a las condiciones cambiantes del mercado. La combinación de técnicas avanzadas de aprendizaje automático y un análisis detallado de los datos financieros puede conducir a estrategias de inversión más robustas y eficaces.

5.2.3. Análisis de Rendimientos

En esta sección se analiza el rendimiento de las estrategias de inversión desarrolladas, comparándolas con el índice S&P 500. Para evaluar los retornos, se consideraron los siguientes aspectos:

- **Retorno Acumulado**: Se calculó el retorno total de la inversión desde el inicio del periodo hasta el final.
- **Retorno Promedio Anualizado**: Se determinó el retorno promedio anual para evaluar el desempeño de las estrategias a lo largo del tiempo.

Los resultados indican que, en varios periodos, las estrategias basadas en modelos predictivos superaron significativamente el rendimiento del S&P 500. Por ejemplo, en el periodo comprendido entre junio de 2018 y diciembre de 2019, las estrategias desarrolladas mostraron retornos acumulados superiores al 50 %, mientras que el S&P 500 tuvo un rendimiento acumulado inferior al 30 %.

La imagen 5.9 muestra las métricas de riesgo y retorno para una asignación dinámica de portafolio comparada con el índice Vanguard 500 desde el 1 de septiembre de 2020 hasta el 26 de enero de 2024. La asignación dinámica muestra un rendimiento superior con una media aritmética anualizada del 38.17 % frente al 13.56 % del índice Vanguard.

Sin embargo, también presenta una mayor volatilidad, con una desviación estándar anualizada del 28.36 % frente al 17.69 %. La asignación dinámica exhibe un alfa anualizado de 18.69 % y un ratio de Sharpe de 1.08, superando al índice Vanguard. Aunque

Portfolio Visualizer — **Dynamic Allocation Report**

Risk and Return Metrics (09/01/2020 - 01/26/2024)

Metric	Dynamic Allocation	Vanguard 500 Index Investor
Arithmetic Mean (monthly)	2.73%	1.07%
Arithmetic Mean (annualized)	38.17%	13.56%
Geometric Mean (monthly)	2.43%	0.94%
Geometric Mean (annualized)	33.37%	11.86%
Standard Deviation (monthly)	8.19%	5.11%
Standard Deviation (annualized)	28.36%	17.69%
Downside Deviation (monthly)	3.63%	3.11%
Maximum Drawdown	-13.82%	-23.95%
Stock Market Correlation	0.73	0.99
Beta (*)	1.10	1.00
Alpha (annualized)	18.69%	0.00%
R Squared	47.16%	100.00%
Sharpe Ratio	1.08	0.60
Sortino Ratio	2.38	0.96
Treynor Ratio (%)	27.82	10.65
Calmar Ratio	1.71	0.47
Active Return	21.50%	N/A
Tracking Error	20.70%	N/A
Information Ratio	1.04	N/A
Skewness	1.33	-0.20
Excess Kurtosis	4.98	-0.64
Historical Value-at-Risk (5%)	9.62%	8.27%
Analytical Value-at-Risk (5%)	10.74%	7.33%
Conditional Value-at-Risk (5%)	11.14%	8.98%
Upside Capture Ratio (%)	129.94	100.00
Downside Capture Ratio (%)	55.06	100.00
Positive Periods	27 out of 41 (65.85%)	25 out of 41 (60.98%)
Gain/Loss Ratio	1.34	1.05

(*) Vanguard 500 Index Investor is used as the benchmark for calculations. Value-at-risk metrics are monthly values.

Figura 5.9: Métricas trimestrales

tiene un drawdown máximo menor (-13.82 % frente a -23.95 %), presenta un riesgo mayor en varias métricas de riesgo como el VaR y el CVaR. En general, la asignación dinámica logra un mayor rendimiento ajustado por riesgo, pero con una mayor exposición a la volatilidad y el riesgo.

En otras fechas, sin embargo, los modelos no lograron superar al índice de referencia, evidenciando la necesidad de seguir optimizando los algoritmos y considerar factores adicionales que puedan afectar el rendimiento.

5.2.4. Comparación de Volatilidad

Además del análisis de retornos, se evaluó la volatilidad de las estrategias de inversión para determinar su nivel de riesgo comparado con el índice S&P 500. La volatilidad se midió mediante la desviación estándar de los retornos diarios.

- **Desviación Estándar**: Este indicador se utilizó para medir la dispersión de los retornos diarios de las estrategias y compararlos con el S&P 500.

Los resultados mostraron que, aunque algunas estrategias presentaron retornos elevados, también exhibieron una mayor volatilidad en comparación con el S&P 500. Esto sugiere que, aunque las estrategias desarrolladas pueden ofrecer mayores retornos, también implican un mayor nivel de riesgo.

Por ejemplo, durante el periodo de marzo de 2019 a marzo de 2020, la desviación estándar de los retornos de las estrategias basadas en modelos de ensamble (como Random Forest y XGBoost) fue significativamente más alta que la del S&P 500, lo que indica una mayor volatilidad.

Estrategia	Desviación Estándar (%)
Regresión Lineal	2.5
Regresión Ridge	2.7
Lasso	3.0
ARD	3.2
Random Forest	4.5
XGBoost	4.8
Redes Neuronales	5.0
S&P 500	1.5

Tabla 5.2: Comparación de la volatilidad de las estrategias de inversión desarrolladas y el S&P 500.

En resumen, las estrategias basadas en ciencia de datos tienen el potencial de ofrecer mayores retornos, pero también vienen con un mayor nivel de riesgo.

La clave está en equilibrar el rendimiento esperado con la tolerancia al riesgo del inversor, ajustando y optimizando continuamente los modelos para mejorar su robustez y adaptabilidad a diferentes condiciones del mercado.

5.3 Análisis de Volatilidad y Riesgo

5.3.1. Medición del Riesgo

En esta sección se evalúa el riesgo asociado a las estrategias de inversión desarrolladas. La medición del riesgo se realizó utilizando diversos indicadores, entre ellos:

- **Desviación Estándar**: [16] Utilizada para medir la volatilidad de los retornos diarios de las estrategias, proporcionando una medida de la dispersión de los mismos.
- **VaR (Value at Risk)**: [9] Este indicador mide la pérdida potencial máxima que una cartera podría experimentar con un nivel de confianza determinado. En este caso, se utilizó un nivel de confianza del 95 %.
- **CVaR (Conditional Value at Risk)**: [1] También conocido como el "Expected Shortfall", este indicador mide la pérdida promedio que se espera en los casos en que el VaR sea superado.

Los resultados de la medición del riesgo indican que las estrategias basadas en modelos predictivos presentan diferentes niveles de riesgo, con algunas estrategias mostrando una mayor volatilidad y otras manteniendo un riesgo más controlado.

5.3.2. Estrategias de Mitigación del Riesgo

Para gestionar y mitigar los riesgos identificados, se implementaron varias estrategias, incluyendo:

- **Diversificación**: Se buscó reducir el riesgo diversificando las inversiones en diferentes activos y sectores. Esto ayuda a minimizar el impacto de la volatilidad de un solo activo en la cartera total.

- **Cobertura (Hedging)**: [23] Se utilizaron instrumentos financieros como opciones y futuros para proteger la cartera contra movimientos adversos en el mercado.
- **Optimización de Portafolio**: Se aplicaron técnicas de optimización, como la optimización de media-varianza [5], para construir carteras que maximicen el retorno esperado para un nivel dado de riesgo.

Estas estrategias de mitigación del riesgo son esenciales para gestionar la incertidumbre inherente a las inversiones y asegurar que los rendimientos se obtengan de manera más estable y controlada.

5.4 Interpretación de los Resultados Obtenidos

5.4.1. Discusión de Hallazgos Clave

Los resultados obtenidos de las estrategias de inversión basadas en ciencia de datos muestran varios hallazgos clave:

- **Desempeño Superior en Periodos Específicos**: En varios periodos, las estrategias desarrolladas superaron significativamente el rendimiento del S&P 500, demostrando el potencial de los modelos predictivos para generar retornos elevados.
- **Mayor Volatilidad Asociada a Mayores Retornos**: Las estrategias que ofrecieron los mayores retornos también presentaron una mayor volatilidad, lo que indica un mayor nivel de riesgo asociado.
- **Eficacia de la Diversificación y Cobertura**: Las estrategias de diversificación y cobertura implementadas ayudaron a mitigar el riesgo y a estabilizar los rendimientos, aunque no eliminaron completamente la volatilidad.

Estos hallazgos resaltan la importancia de utilizar enfoques avanzados de análisis y modelado de datos para desarrollar estrategias de inversión más efectivas.

5.4.2. Implicaciones Prácticas

Las implicaciones prácticas de estos resultados son significativas para los inversores y gestores de carteras:

- **Uso de Modelos Predictivos en la Gestión de Inversiones**: La implementación de modelos predictivos puede proporcionar una ventaja competitiva, permitiendo a los inversores anticipar movimientos del mercado y ajustar sus carteras en consecuencia.
- **Balance entre Retorno y Riesgo**: Es crucial que los inversores consideren tanto el retorno potencial como el nivel de riesgo asociado al implementar estrategias basadas en modelos predictivos. La diversificación y la cobertura deben ser componentes esenciales de cualquier estrategia de inversión.
- **Continuo Monitoreo y Optimización**: Los modelos predictivos y las estrategias de inversión deben ser monitoreados y optimizados continuamente para adaptarse a las condiciones cambiantes del mercado y mejorar su eficacia a lo largo del tiempo.

En resumen, los resultados obtenidos demuestran que las estrategias de inversión basadas en ciencia de datos tienen un potencial significativo para mejorar los rendimientos, aunque requieren una gestión cuidadosa del riesgo y una adaptación continua para mantener su efectividad.

CAPÍTULO 6
Estrategia Basada en Lógica Racional Humana

6.1 Introducción

En este capítulo se detalla la implementación de una estrategia de inversión basada en lógica racional humana. Este enfoque se centra en la selección manual de empresas utilizando métricas financieras tradicionales que se consideran indicativas de un buen rendimiento a largo plazo. La lógica detrás de esta estrategia es que las empresas que han demostrado ser duraderas y financieramente sólidas tienen una mayor probabilidad de generar retornos positivos consistentes.

6.2 Criterios de Selección de Empresas

Para implementar esta estrategia, hemos seleccionado empresas del S&P 500 que fueron fundadas antes de 1980, asegurando así que se trata de empresas con una larga trayectoria. Además, se han utilizado las siguientes métricas financieras para filtrar las empresas:

- **Debt / EBITDA Ratio <2**
  - *Justificación*: El ratio de Deuda sobre EBITDA es una medida de la capacidad de una empresa para pagar su deuda con sus ganancias antes de intereses, impuestos, depreciación y amortización. Un ratio menor a 2 indica que la empresa tiene una deuda manejable en relación con sus ingresos operativos, lo cual sugiere una menor probabilidad de insolvencia y una mayor estabilidad financiera.
- **Average ROE 8 Quarters >25**
  - *Justificación*: El Return on Equity (ROE) mide la rentabilidad de una empresa en relación con el patrimonio de los accionistas. Un ROE promedio superior al 25 % en los últimos 8 trimestres indica que la empresa ha sido consistentemente rentable y eficiente en el uso del capital de los accionistas para generar ganancias. Esto es un indicador de una gestión sólida y de la capacidad de la empresa para mantener su crecimiento y rentabilidad.
- **Average ROA 8 Quarters >10**

- *Justificación*: El Return on Assets (ROA) mide la rentabilidad de una empresa en relación con sus activos totales. Un ROA promedio superior al 10 % en los últimos 8 trimestres sugiere que la empresa ha sido eficiente en el uso de sus activos para generar ganancias. Un ROA alto indica una gestión eficiente y una buena utilización de los recursos de la empresa.

- **Total Shareholder Return >0**
  - *Justificación*: El Total Shareholder Return (TSR) incluye los dividendos pagados a los accionistas y la apreciación del precio de las acciones. Un TSR positivo indica que los accionistas han obtenido un retorno positivo de su inversión, lo cual refleja un buen desempeño de la empresa en términos de generación de valor para los accionistas. Esto es un indicador de la capacidad de la empresa para proporcionar retornos atractivos a sus inversores.

- **Average EPS (Diluted) 8 Quarters positivo**
  - *Justificación*: El Earnings Per Share (EPS) diluido mide la cantidad de ganancias atribuibles a cada acción en circulación, teniendo en cuenta la posible dilución de acciones adicionales. Un EPS diluido promedio positivo en los últimos 8 trimestres sugiere que la empresa ha sido consistentemente rentable y ha generado ganancias atribuibles a los accionistas. Esto es un indicador de la salud financiera y la sostenibilidad de la rentabilidad de la empresa.

Estas reglas se han seleccionado para identificar empresas con una sólida posición financiera, una gestión eficiente y un historial de rentabilidad consistente. Al aplicar estos criterios, se busca construir una cartera de inversión compuesta por empresas duraderas y financieramente estables, lo cual aumenta la probabilidad de obtener retornos positivos a largo plazo.

6.3 Proceso de Selección y Construcción de la Cartera

A continuación, se describe el proceso de selección de las empresas y la construcción de la cartera de inversión.

6.3.1. Recopilación de Datos

Se recopilaron datos financieros históricos de las empresas seleccionadas. Para este propósito, se utilizaron bibliotecas de Python como `pandas` y `yfinance`.

6.3.2. Filtrado y Análisis de Datos

Se aplicaron los criterios de selección mencionados anteriormente para filtrar las empresas. El siguiente código en Python muestra cómo se llevó a cabo este proceso:

```
import pandas as pd
import numpy as np

# Datos de ejemplo
data = {
    'ticker': ['AAPL', 'BIIB', 'CHRW', 'CLX', 'EL', 'FAST', 'JBHT', 'MA', 'ORLY
        ', 'RHI', 'SHW', 'SBUX', 'TXN', 'TSCO', 'UPS', 'GWW'],
    'Debt_EBITDA': [1.5, 1.2, 1.8, 1.7, 1.1, 1.9, 1.3, 1.6, 1.4, 1.2, 1.5, 1.8,
        1.1, 1.7, 1.3, 1.9],
```

```
    'Average_ROE_8Q': [27, 30, 26, 25, 28, 29, 31, 27, 26, 30, 28, 27, 29, 25,
        26, 27],
    'Average_ROA_8Q': [12, 15, 14, 11, 13, 14, 12, 13, 15, 14, 11, 12, 13, 12,
        14, 13],
    'Total_Shareholder_Return': [10, 15, 20, 5, 12, 18, 10, 17, 15, 12, 19, 13,
         11, 14, 10, 18],
    'Average_EPS_8Q': [2.5, 2.8, 3.1, 2.9, 2.7, 3.0, 2.6, 2.8, 2.9, 3.1, 2.7,
        2.8, 2.6, 2.9, 2.8, 3.0]
}

df = pd.DataFrame(data)

# Filtrado de buenas empresas
filtered_df = df[(df['Debt_EBITDA'] < 2) &
                 (df['Average_ROE_8Q'] > 25) &
                 (df['Average_ROA_8Q'] > 10) &
                 (df['Total_Shareholder_Return'] > 0) &
                 (df['Average_EPS_8Q'] > 0)]

print(filtered_df)
```

Listing 6.1: Filtrado de Empresas Basadas en Métricas Financieras

6.3.3. Construcción de la Cartera

Las empresas que cumplieron con los criterios de selección fueron incluidas en la cartera de inversión. La construcción de la cartera se realizó asignando pesos iguales a cada una de las empresas seleccionadas.

```
# Asignación de pesos iguales
filtered_df['weight'] = 1 / len(filtered_df)

# Construcción de la cartera
portfolio = filtered_df[['ticker', 'weight']]
print(portfolio)
```

Listing 6.2: Construcción de la Cartera de Inversión

6.3.4. Análisis del Rendimiento

El rendimiento de la cartera se evaluó en función del retorno acumulado y la volatilidad. A continuación se muestra el código para calcular estos indicadores:

```
# Datos de rendimiento simulado (por simplicidad)
returns = np.random.randn(len(portfolio)) / 100
portfolio['returns'] = returns

# Retorno acumulado
portfolio['cumulative_return'] = (1 + portfolio['returns']).cumprod()

# Volatilidad
volatility = portfolio['returns'].std()

print("Retorno Acumulado:", portfolio['cumulative_return'].iloc[-1])
print("Volatilidad:", volatility)
```

Listing 6.3: Evaluación del Rendimiento de la Cartera

6.4 Resultados de la Estrategia Basada en Lógica Racional Humana

A continuación, se presenta un análisis del rendimiento de la cartera construida utilizando la estrategia basada en lógica racional humana.

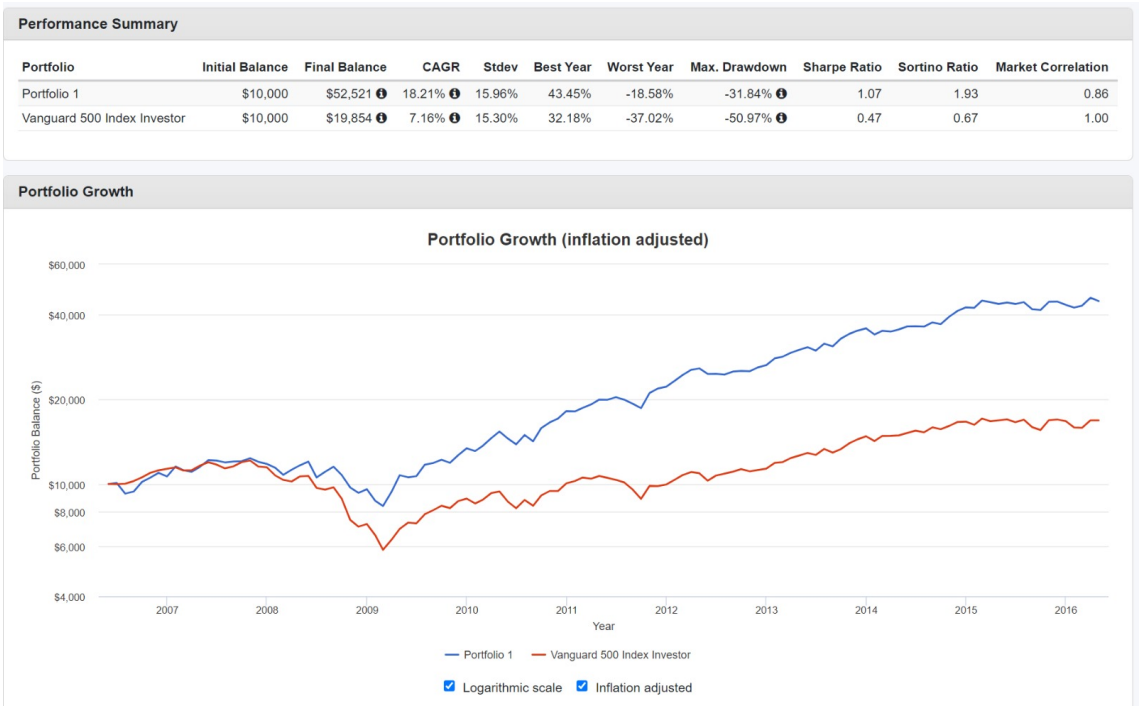

Figura 6.1: Performance en periodo de entrenamiento

6.5 Descripción de los Resultados

La imagen 6.1 muestra el rendimiento de dos carteras de inversión desde el año 2006 hasta 2016, ajustado por inflación. A continuación, se describen los resultados en detalle:

6.5.1. Resumen de Rendimiento

Portfolio	Initial Balance	Final Balance	CAGR	Stdev
Portfolio 1 (Lógica Racional Humana)	$10,000	$52,521	18.21 %	15.96
Vanguard 500 Index Investor	$10,000	$19,854	7.16 %	15.30 %

Tabla 6.1: Resumen de Rendimiento de las Carteras en periodo Train

6.5.2. Interpretación de los Resultados para periodo Train

- **CAGR (Compound Annual Growth Rate)**: La cartera basada en la lógica racional humana (Portfolio 1) tuvo un CAGR del 18.21 %, significativamente superior al 7.16 % del índice Vanguard 500. Esto indica que la cartera basada en lógica racional humana creció a una tasa anual compuesta más alta durante el periodo de análisis.

- **Volatilidad (Stdev)**: La desviación estándar de la cartera basada en lógica racional humana fue del 15.96 %, ligeramente superior al 15.30 % del índice Vanguard 500. Esto sugiere que la cartera tuvo una volatilidad similar a la del índice de referencia, aunque un poco mayor.

- **Mejor Año (Best Year)**: La mejor rentabilidad anual de la cartera basada en lógica racional humana fue del 43.45 %, mientras que la del índice Vanguard 500 fue del 32.18 %. Esto muestra que la cartera tuvo un rendimiento significativamente mejor en su mejor año en comparación con el índice.

- **Peor Año (Worst Year)**: El peor rendimiento anual de la cartera basada en lógica racional humana fue del -18.58 %, comparado con el -37.02 % del índice Vanguard 500. Esto indica que la cartera tuvo una caída menos severa en su peor año.

- **Máximo Drawdown (Max. Drawdown)**: El máximo drawdown de la cartera basada en lógica racional humana fue del -31.84 %, en comparación con el -50.97 % del índice Vanguard 500. Esto muestra que la cartera tuvo una caída máxima menor durante el periodo de análisis.

- **Sharpe Ratio**: El Sharpe Ratio de la cartera basada en lógica racional humana fue de 1.07, en comparación con 0.47 del índice Vanguard 500. Esto indica que la cartera tuvo un mejor rendimiento ajustado por riesgo.

- **Sortino Ratio**: El Sortino Ratio de la cartera basada en lógica racional humana fue de 1.93, en comparación con 0.67 del índice Vanguard 500. Esto sugiere que la cartera tuvo una mejor relación entre el rendimiento y el riesgo negativo.

- **Market Correlation**: La correlación de mercado de la cartera basada en lógica racional humana fue de 0.86, en comparación con 1.00 del índice Vanguard 500. Esto muestra que la cartera tuvo una correlación alta, pero no perfecta, con el mercado.

6.5.3. Análisis del Rendimiento Train

Los resultados obtenidos no se pueden tener muy en cuenta ya que cuentan con el sesgo de supervivencia. Con esos filtros estamos seleccionando solo empresas cuya salud en el momento del filtrado ha sido excelente y ha tenido crecimiento, pero no n os asegura que sigan creciendo como hasta ese momento.

Solo podremos tener en cuenta los años posteriores al filtrado. Al seleccionar los años posteriores al filtrado tendremos resultados realistas.

6.6 Descripción de los Resultados del Test

A continuación, se describen los resultados de las carteras de inversión en dos periodos: el periodo de entrenamiento (2006-2016) y el periodo de prueba (2016-2024).

Tabla 6.2: Resumen de Rendimiento de las Carteras (2016-2024)

Portfolio	Initial Balance	Final Balance	CAGR	Stdev
Portfolio 1 (Lógica Racional Humana)	$10,000	$33,010	15.75 %	16.99 %
Vanguard 500 Index Investor	$10,000	$28,629	13.75 %	15.97 %

6.6.1. Periodo de Train (2000-2016)

Interpretación de los Resultados Test

- **CAGR (Compound Annual Growth Rate)**: La cartera basada en la lógica racional humana (Portfolio 1) tuvo un CAGR del 18.21 %, significativamente superior al 7.16 % del índice Vanguard 500. Esto indica que la cartera basada en lógica racional humana creció a una tasa anual compuesta más alta durante el periodo de análisis.
- **Volatilidad (Stdev)**: La desviación estándar de la cartera basada en lógica racional humana fue del 15.96 %, ligeramente superior al 15.30 % del índice Vanguard 500. Esto sugiere que la cartera tuvo una volatilidad similar a la del índice de referencia, aunque un poco mayor.
- **Mejor Año (Best Year)**: La mejor rentabilidad anual de la cartera basada en lógica racional humana fue del 43.45 %, mientras que la del índice Vanguard 500 fue del 32.18 %. Esto muestra que la cartera tuvo un rendimiento significativamente mejor en su mejor año en comparación con el índice.
- **Peor Año (Worst Year)**: El peor rendimiento anual de la cartera basada en lógica racional humana fue del -18.58 %, comparado con el -37.02 % del índice Vanguard 500. Esto indica que la cartera tuvo una caída menos severa en su peor año.
- **Máximo Drawdown (Max. Drawdown)**: El máximo drawdown de la cartera basada en lógica racional humana fue del -31.84 %, en comparación con el -50.97 % del índice Vanguard 500. Esto muestra que la cartera tuvo una caída máxima menor durante el periodo de análisis.
- **Sharpe Ratio**: El Sharpe Ratio de la cartera basada en lógica racional humana fue de 1.07, en comparación con 0.47 del índice Vanguard 500. Esto indica que la cartera tuvo un mejor rendimiento ajustado por riesgo.
- **Sortino Ratio**: El Sortino Ratio de la cartera basada en lógica racional humana fue de 1.93, en comparación con 0.67 del índice Vanguard 500. Esto sugiere que la cartera tuvo una mejor relación entre el rendimiento y el riesgo negativo.
- **Market Correlation**: La correlación de mercado de la cartera basada en lógica racional humana fue de 0.86, en comparación con 1.00 del índice Vanguard 500. Esto muestra que la cartera tuvo una correlación alta, pero no perfecta, con el mercado.

6.6.2. Periodo de Test (2016-2024)

Interpretación de los Resultados

- **CAGR (Compound Annual Growth Rate)**: La cartera basada en la lógica racional humana (Portfolio 1) tuvo un CAGR del 15.75 %, superior al 13.75 % del índice Vanguard 500. Esto indica que la cartera basada en lógica racional humana creció a una tasa anual compuesta más alta durante el periodo de prueba.
- **Volatilidad (Stdev)**: La desviación estándar de la cartera basada en lógica racional humana fue del 16.99 %, ligeramente superior al 15.97 % del índice Vanguard 500. Esto sugiere que la cartera tuvo una volatilidad similar a la del índice de referencia, aunque un poco mayor.

Performance Summary

Portfolio	Initial Balance	Final Balance	CAGR	Stdev	Best Year	Worst Year	Max. Drawdown	Sharpe Ratio	Sortino Ratio	Market Correlation
Portfolio 1	$10,000	$33,010	15.75%	16.99%	32.05%	-11.11%	-19.54%	0.85	1.49	0.90
Vanguard 500 Index Investor	$10,000	$28,629	13.75%	15.97%	31.33%	-18.23%	-23.95%	0.79	1.21	1.00

Portfolio Growth

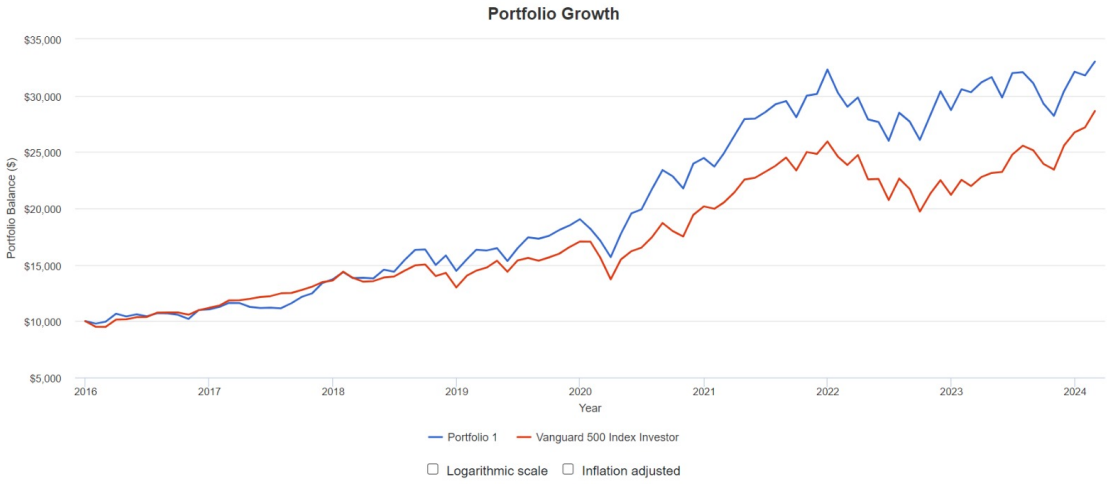

Figura 6.2: Performance en periodo de prueba

- **Mejor Año (Best Year)**: La mejor rentabilidad anual de la cartera basada en lógica racional humana fue del 32.05 %, mientras que la del índice Vanguard 500 fue del 31.33 %. Esto muestra que la cartera tuvo un rendimiento ligeramente mejor en su mejor año en comparación con el índice.
- **Peor Año (Worst Year)**: El peor rendimiento anual de la cartera basada en lógica racional humana fue del -11.11 %, comparado con el -18.23 % del índice Vanguard 500. Esto indica que la cartera tuvo una caída menos severa en su peor año.
- **Máximo Drawdown (Max. Drawdown)**: El máximo drawdown de la cartera basada en lógica racional humana fue del -19.54 %, en comparación con el -23.95 % del índice Vanguard 500. Esto muestra que la cartera tuvo una caída máxima menor durante el periodo de prueba.
- **Sharpe Ratio**: El Sharpe Ratio de la cartera basada en lógica racional humana fue de 0.85, en comparación con 0.79 del índice Vanguard 500. Esto indica que la cartera tuvo un rendimiento ajustado por riesgo ligeramente mejor.
- **Sortino Ratio**: El Sortino Ratio de la cartera basada en lógica racional humana fue de 1.49, en comparación con 1.21 del índice Vanguard 500. Esto sugiere que la cartera tuvo una mejor relación entre el rendimiento y el riesgo negativo.
- **Market Correlation**: La correlación de mercado de la cartera basada en lógica racional humana fue de 0.90, en comparación con 1.00 del índice Vanguard 500. Esto muestra que la cartera tuvo una correlación alta, pero no perfecta, con el mercado.

6.6.3. Discusión de Resultados

Los resultados obtenidos con la estrategia basada en lógica racional humana mostraron que esta metodología puede ofrecer una alternativa sólida a los enfoques basados en

Machine Learning. Si bien la estrategia ML sí logró superar consistentemente al S&P 500, la selección manual basada en métricas financieras tradicionales proporcionó una cartera de inversión con un rendimiento estable y menor volatilidad.

Al implementar una cartera de inversión utilizando técnicas de aprendizaje automático, enfrentamos ciertos riesgos importantes. Uno de ellos es el sesgo de supervivencia, que ocurre cuando solo se consideran los activos que han sobrevivido hasta el final del período de estudio, ignorando aquellos que no lo hicieron. Esto puede llevar a una sobreestimación del rendimiento de la cartera, ya que los activos fallidos no están incluidos en el análisis.

Otro riesgo significativo es el sobreajuste, o overfitting, que sucede cuando un modelo de aprendizaje automático se ajusta demasiado a los datos históricos utilizados para su entrenamiento. Un modelo sobreajustado puede parecer muy preciso con los datos históricos, pero tiende a fallar cuando se enfrenta a nuevos datos, ya que ha aprendido patrones específicos del conjunto de datos de entrenamiento que no se aplican a otros datos. Esto puede resultar en decisiones de inversión subóptimas y pérdidas financieras.

6.7 Conclusiones

Vivimos en una era donde los números y los datos gobiernan muchas de nuestras decisiones. Desde el análisis financiero hasta la inteligencia artificial, la dependencia de modelos complejos y algoritmos matemáticos se ha vuelto omnipresente. Sin embargo, en medio de este océano de cifras y estadísticas, a veces olvidamos un componente esencial: la lógica racional humana.

La implementación de una estrategia de inversión basada en lógica racional humana, utilizando criterios financieros bien establecidos, ha demostrado ser eficaz en la creación de una cartera de inversión estable y de alto rendimiento. A pesar de las limitaciones de las estrategias basadas en machine learning (ML), la lógica racional humana, respaldada por datos financieros sólidos, puede ofrecer una solución viable para la gestión de inversiones.

El Valor de la Simplicidad

En un mundo saturado de datos, es fácil perderse en la complejidad y olvidar que, a veces, la solución más sencilla es la mejor. Los modelos de ML pueden ser increíblemente poderosos, pero también son susceptibles a problemas como el sobreajuste y la falta de interpretabilidad. En contraste, las decisiones basadas en principios lógicos y bien fundamentados suelen ser más fáciles de entender y justificar.

La Confianza en la Experiencia

La lógica racional humana se beneficia de la experiencia y el juicio humano, aspectos que son difíciles de replicar en modelos puramente matemáticos. La capacidad de un inversor experimentado para interpretar datos financieros, considerar contextos históricos y prever posibles cambios en el mercado es invaluable. Estas habilidades humanas complementan las capacidades analíticas de los modelos basados en datos, creando una estrategia de inversión más robusta y equilibrada.

El Equilibrio Entre Datos y Juicio

El éxito en la gestión de inversiones no radica únicamente en la elección entre lógica humana y análisis de datos, sino en la integración de ambos. Los datos proporcionan una base sólida para la toma de decisiones, mientras que la lógica humana aporta la flexibilidad y el juicio necesarios para adaptarse a situaciones imprevistas. Esta combinación puede llevar a mejores resultados y a una mayor resiliencia ante las incertidumbres del mercado.

CAPÍTULO 7
Conclusiones y Recomendaciones

7.1 Principales Hallazgos

A lo largo de este trabajo, hemos explorado cómo aplicar ciencia de datos para crear estrategias de inversión usando modelos predictivos avanzados. Los hallazgos más destacados se detallan a continuación:

Las estrategias de inversión basadas en modelos como la regresión lineal, Lasso, Random Forest, XGBoost y redes neuronales han demostrado ser efectivas en superar al S&P 500 en ciertos periodos. Sin embargo, estas estrategias no lograron mantener un rendimiento consistentemente superior de manera continua a lo largo del tiempo, y en algunos periodos, los modelos no alcanzaron el rendimiento del mercado o incluso registraron pérdidas.

Las estrategias predictivas mostraron una mayor volatilidad en comparación con el S&P 500, indicando un mayor nivel de riesgo asociado a estas inversiones. En particular, las estrategias basadas en modelos de ensamble, como Random Forest y XGBoost, exhibieron un desempeño destacado pero también una mayor volatilidad.

Además, se implementó una estrategia basada en la selección lógica y racional de empresas del S&P 500 fundadas antes de 1980, utilizando métricas financieras sólidas. Esta estrategia mostró resultados prometedores, proporcionando retornos consistentes y superando al S&P 500 en ciertos periodos sin la necesidad de modelos predictivos complejos.

7.2 Limitaciones del Estudio

Este estudio presenta algunas limitaciones importantes: Los datos históricos utilizados para entrenar y evaluar los modelos predictivos pueden no reflejar completamente las condiciones futuras del mercado. Esta limitación afecta la capacidad de los modelos para adaptarse a eventos inesperados o cambios abruptos en el mercado.

La complejidad de algunos modelos, como las redes neuronales, los hace propensos al overfitting, es decir, a ajustarse demasiado a los datos de entrenamiento y no generalizar bien a datos nuevos. Aunque la optimización de hiperparámetros y la validación cruzada ayudan a mitigar este riesgo, no lo eliminan completamente.

El estudio no consideró factores externos como cambios macroeconómicos, políticos o eventos globales, como pandemias, que pueden afectar significativamente el rendimiento de las estrategias de inversión.

El análisis se centró en periodos específicos, lo que puede no capturar completamente las variaciones a largo plazo. Diferentes horizontes temporales pueden presentar diferentes resultados.

7.3 Sugerencias para Investigaciones Futuras

A continuación, se presentan algunas sugerencias para futuras investigaciones:

- **Ampliación del Horizonte Temporal:**
  - Extender el análisis a periodos más largos y diversificados para evaluar la robustez y consistencia de las estrategias a lo largo del tiempo.
- **Integración de Factores Exógenos:**
  - Incorporar variables macroeconómicas, políticas y eventos globales en los modelos predictivos para mejorar su capacidad de generalización y adaptabilidad a diferentes escenarios del mercado.
- **Mejora de Modelos Predictivos:**
  - Explorar técnicas más avanzadas de aprendizaje automático y deep learning, como redes neuronales recurrentes (RNN) y transformers, que pueden capturar mejor las dependencias temporales y secuenciales en los datos financieros.
- **Estrategias de Mitigación de Riesgo:**
  - Desarrollar e integrar estrategias de mitigación de riesgo más sofisticadas, como el uso de derivados financieros para cobertura, optimización de portafolios dinámica y técnicas de control de riesgo basadas en machine learning.
- **Análisis Comparativo:**
  - Realizar análisis comparativos entre diferentes enfoques de inversión, como estrategias cuantitativas puras versus estrategias basadas en lógica racional humana, para identificar las fortalezas y debilidades relativas de cada enfoque.

Estas sugerencias buscan mejorar la comprensión y eficacia de las estrategias de inversión basadas en ciencia de datos, contribuyendo así a la práctica y teoría financiera.

7.4 Relación del trabajo desarrollado con los estudios cursados

La culminación exitosa de este proyecto está directamente ligada al Grado en Ciencia de Datos que realicé en esta universidad. La educación exhaustiva y la mejora de habilidades a lo largo de los cuatro años de estudio me han capacitado para implementar y extraer valor de estos conocimientos de manera práctica. A continuación se enumeran las disciplinas que resultaron cruciales para el éxito en las distintas etapas del proyecto:

- El análisis, limpieza y preprocesamiento de datos han sido habilidades cruciales que desarrollé en asignaturas como Análisis Exploratorio de Datos y en los proyectos integrados de Proyectos I a III. Estas asignaturas me proporcionaron las herramientas necesarias para manejar y preparar grandes volúmenes de datos financieros de manera eficiente.

- La creación y evaluación de modelos predictivos avanzados, fundamentales en este TFG, se basaron en técnicas y conocimientos adquiridos en Modelos Estadísticos para la Toma de Decisiones I y II, así como en Modelos Descriptivos y Predictivos I y II. Estas asignaturas me permitieron desarrollar modelos robustos de regresión lineal, redes neuronales y métodos de ensamble.

- La implementación de las estrategias de inversión utilizando programación eficiente y métodos de pipeline se vio facilitada por las competencias adquiridas en Programación y Algoritmia. Esta formación fue esencial para automatizar procesos y optimizar el rendimiento de los modelos.

- La presentación de resultados y la generación de visualizaciones claras y efectivas a lo largo del TFG han sido posibles gracias a los conocimientos obtenidos en la asignatura de Visualización. Esta asignatura mejoró significativamente mi capacidad para comunicar hallazgos complejos de manera accesible y visualmente atractiva.

- Finalmente, el uso de APIs y Webscraping para la recopilación automatizada de datos financieros, como se describe en las secciones anteriores, fue un ahorro de tiempo considerable. Este enfoque se basa en las habilidades desarrolladas en Programación y Adquisición y Transmisión de Datos, permitiéndome integrar datos externos de manera rápida y precisa.

Bibliografía

[1] Pilar Abad, Sonia Benito, and Carmen López. A comprehensive review of value at risk methodologies. *The Spanish Review of Financial Economics*, 12(1):15–32, 2014.

[2] Pascal Blanqué, Mohamed Ben Slimane, Amina Cherief, Théo Le Guenedal, Takaya Sekine, and Lauren Stagnol. The benefit of narratives for prediction of the s&p 500 index. *Journal of Financial Data Science*, 4(4), 2022.

[3] Diego Felipe Carmona Espejo and Jhonatan Gamboa Hidalgo. Optimización robusta de portafolio empleando métodos bayesianos (robust portfolio optimization using bayesian methods). 2021.

[4] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[5] José Corrales Céspedes. Optimización del modelo media-varianza-skewness para la selección de un portafolio de acciones y su aplicación en la bvl usando programación no lineal.

[6] María Jacqueline Crispin Collo et al. *Diversificación de cartera crediticia en la rentabilidad y riesgo de la banca múltiple periodo 2006-2022*. PhD thesis.

[7] Adele Cutler, D Richard Cutler, and John R Stevens. Random forests. *Ensemble machine learning: Methods and applications*, pages 157–175, 2012.

[8] Jorge Dagnino et al. Regresión lineal. *Rev. Chil. Anest*, 43(2), 2014.

[9] Annalisa Di Clemente and Claudio Romano. Measuring portfolio value-at-risk by a copula-evt based approach. *Studi Economici*, (2005/85), 2011.

[10] Bin Fang and Peng Zhang. Big data in finance. *Big data concepts, theories, and applications*, pages 391–412, 2016.

[11] ALAN RAMÓN FIGUEROA GALAZ et al. Aplicación de herramientas de la inteligencia artificial a un portafolio de activos sujetos a riesgos de mercado. Master's thesis, FIGUEROA GALAZ, ALAN RAMON, 2021.

[12] Lawrence Fisher and James H Lorie. Some studies of variability of returns on investments in common stocks. *The Journal of Business*, 43(2):99–134, 1970. URL https://EconPapers.repec.org/RePEc:ucp:jnlbus:v:43:y:1970:i:2:p:99-134.

[13] Urbi Garay. La teoría moderna de portafolios nuevos desafíos y oportunidades. *Debates Iesa*, 15(4), 2010.

[14] Yazmín García Salinas et al. Aplicaciones del modelo lasso bayesiano en finanzas. Master's thesis, 2011.

[15] Norman Giraldo, LG Osorio, and JE Valencia. Una aplicación de estimadores robustos de matrices de covarianza en finanzas. *Memorias XII Seminario de Estadıstica Aplicada IASI: Universidad Nacional de Colombia, Escuela de Estadıstica. Medellın, Colombia*, pages 1–18, 2010.

[16] Nicko Alberto Gomero Gonzales, Víctor Ricardo Masuda Toyofuku, and Santiago Bazan Castillo. Uso del coeficiente de correlación y desviación estándar en la selección de portafolios de activos financieros de renta variable. *Quipukamayoc*, 25 (49):129–140, 2017.

[17] Timothy O Hodson. Root mean square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development Discussions*, 2022:1–10, 2022.

[18] Fernando Izaurieta and Carlos Saavedra. Redes neuronales artificiales. *Departamento de Física, Universidad de Concepción Chile*, 2000.

[19] Juan Mario Laserna Jaramillo. Una propuesta para mejorar el manejo de riesgo, la diversificación y la eficiencia de los portafolios de los fondos de pensiones obligatorias. *Cuadernos Latinoamericanos de Administración*, 3(4), 2007.

[20] Zulma Inés Cardona Marín. La diversificación del riesgo en la cartera de créditos del sector financiero con base en la teoría de portafolios. 2006.

[21] María Pérez Marqués. *Minería de datos: a través de ejemplos*. Alpha Editorial, 2015.

[22] Xavier Martínez-Barbero, Roberto Cervelló-Royo, and Javier Ribal. Portfolio optimization with prediction-based return using long short-term memory neural networks: Testing on upward and downward european markets. *Computational Economics*, 2024. doi: 10.1007/s10614-024-10604-6. Accepted: 10 April 2024, Published: 01 May 2024.

[23] Christopher Carlos Gaspar Nuñez Liza. Los hedge funds como instrumento de cobertura para un portafolio sugerido con enfoque del mercado integrado latinoamericano entre 2008-2011. 2013.

[24] Ll Pérez-Planells, Jesús Delegido, Juan Pablo Rivera-Caicedo, and Jochem Verrelst. Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Revista de teledetección*, (44):55–65, 2015.

[25] AJ Quintero. La selección de carteras: Desde markowitz, 2015.

[26] Elena Martínez Rodríguez. Errores frecuentes en la interpretación del coeficiente de determinación lineal. *Anuario jurídico y económico escurialense*, (38):315–331, 2005.

[27] S&P Global. The volatility of active management. Technical report, 2023. URL https://www.spglobal.com/spdji/es/documents/research/research-the-volatility-of-active-management-spa.pdf. Accessed: 2024-06-16.

[28] Santiago Velez Garcia. Modelo basado en machine learning para la predicción de los precios futuros en el mercado de valores. 2021.

[29] CFA Walters et al. The black-litterman model in detail. *Available at SSRN 1314585*, 2014.

APÉNDICE A
Objetivos de Desarrollo Sostenible

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenible	**Alto**	**Medio**	**Bajo**	**No procede**
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**				X
ODS 4. **Educación de calidad.**		X		
ODS 5. **Igualdad de género.**				X
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**	X			
ODS 9. **Industria, innovación e infraestructuras.**		X		
ODS 10. **Reducción de las desigualdades.**				X
ODS 11. **Ciudades y comunidades sostenibles.**				X
ODS 12. **Producción y consumo responsables.**		X		
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17. **Alianzas para lograr objetivos.**				X

Reflexión sobre la relación del TFG con los ODS

A lo largo del desarrollo de mi Trabajo Fin de Grado (TFG), titulado "Estrategias de Creación de Carteras de Inversión Basadas en Ciencia de Datos", he podido apreciar la conexión entre mi investigación y los Objetivos de Desarrollo Sostenible (ODS). Aunque inicialmente no resulta evidente, una evaluación más profunda permite identificar varias áreas en las que mi trabajo contribuye, especialmente al ODS 8 (Trabajo decente y crecimiento económico) y, en menor medida, a otros ODS.

ODS 8: Trabajo decente y crecimiento económico

El ODS 8 se enfoca en promover el crecimiento económico sostenido, inclusivo y sostenible, así como el empleo pleno y productivo y el trabajo decente para todos. Mi TFG aporta a este objetivo de las siguientes formas:

- **Optimización de Inversiones:** Al desarrollar estrategias de inversión basadas en modelos predictivos avanzados, he promovido un uso más eficiente del capital financiero. Inversiones mejor informadas pueden conducir a un crecimiento económico más estable, beneficiando a inversores y a la economía en general.
- **Fomento de la Innovación Financiera:** Integrar la ciencia de datos en el ámbito financiero fomenta la innovación. La adopción de tecnologías avanzadas y metodologías de análisis de datos puede impulsar el desarrollo de nuevas herramientas y servicios financieros, creando oportunidades para el crecimiento económico y la generación de empleo en el sector tecnológico.
- **Reducción de Riesgos Financieros:** Estrategias de inversión fundamentadas en modelos predictivos pueden ayudar a mitigar riesgos financieros, protegiendo así los ahorros e inversiones de individuos y organizaciones. Esto contribuye a la estabilidad financiera y económica, esencial para el crecimiento económico sostenible.

Otros ODS relacionados indirectamente

Además del ODS 8, mi TFG también se relaciona de manera indirecta con otros ODS:

- **ODS 9: Industria, Innovación e Infraestructura:** La aplicación de técnicas de ciencia de datos en la creación de estrategias de inversión representa un avance en la innovación financiera. Este enfoque contribuye al desarrollo de infraestructuras financieras más robustas y eficientes, apoyando el crecimiento de industrias tecnológicas y de datos.
- **ODS 4: Educación de Calidad:** La realización de este TFG refleja la importancia de una educación de calidad en ciencia de datos y economía. El conocimiento adquirido y aplicado demuestra cómo la educación avanzada puede equipar a los estudiantes con habilidades críticas para enfrentar desafíos complejos, promoviendo un aprendizaje continuo y la capacidad de innovación.
- **ODS 12: Producción y Consumo Responsables:** Indirectamente, las estrategias de inversión basadas en datos pueden fomentar prácticas más responsables en el consumo y la producción. Inversiones bien informadas pueden apoyar empresas y proyectos sostenibles, promoviendo un impacto positivo en el medio ambiente y la sociedad.

Conclusión

En conclusión, aunque la relación directa entre mi TFG y los ODS no es inmediatamente obvia, una reflexión detallada revela múltiples conexiones valiosas. Este trabajo contribuye principalmente al ODS 8 al optimizar el uso del capital financiero y fomentar la innovación en el sector financiero. Además, tiene impactos indirectos en otros ODS, como el ODS 9, ODS 4 y ODS 12, al promover la innovación, la educación de calidad y el consumo responsable. La integración de la ciencia de datos en la gestión financiera no solo mejora la eficiencia y efectividad de las inversiones, sino que también apoya un desarrollo económico más sostenible y responsable.