Facultad de Ciencias Económicas y Empresariales
ICADE

COMPARACIÓN ENTRE EL MODELO DE CINCO FACTORES DE FAMA-FRENCH Y UN MODELO DE MACHINE LEARNING EN EL MERCADO EUROPEO

Autor: Juan Bertrán Hervella
Director: Yannis Paraskevopoulos

MADRID | 2025

Índice

Índice de Abreviaturas

CAPM – Capital Asset Pricing Model

OLS – Ordinary Least Squares

ML – Machine Learning

SMB – Small Minus Big

HML – High Minus Low

RMW – Robust Minus Weak

CMA – Conservative Minus Aggressive

RF – Risk-Free Rate

Mkt–RF – Factor de mercado menos tasa libre de riesgo

R^2 – Coeficiente de determinación

MSE – Error cuadrático medio

TFG – Trabajo de Fin de Grado

Resumen y Palabras Clave

Este Trabajo de Fin de Grado analiza y compara el rendimiento del modelo de cinco factores de Fama y French con un modelo de Machine Learning Random Forest Regressor, aplicado a un portafolio construido con acciones europeas de gran capitalización. El estudio utiliza datos mensuales entre 2002 y 2025 integrando los factores del modelo Kenneth French y un portafolio diversificado elaborado mediante datos de mercado obtenidos con yfinance.

La metodología combina herramientas clásicas de econometría como la regresión OLS con técnicas de predicción de Machine Learning, evaluando ambos modelos bajo el mismo marco temporal, la misma muestra y métricas comparables. Los resultados muestran que el modelo lineal presenta un mayor poder explicativo y predictivo, con un R^2 más alto y un MSE más bajo en la prueba. Además, el modelo Random Forest no consigue mejorar las predicciones del modelo lineal, lo que sugiere que la relación entre los factores de Fama-French y los retornos del portafolio europeo es fundamentalmente lineal.

La investigación confirma la robustez del modelo de cinco factores en el contexto europeo y muestra que, en ausencia de un gran volumen de predictores o de fuertes no linealidades, las técnicas de Machine Learning no superan a los modelos teóricos tradicionales. Por ello, este TFG aporta evidencias sobre la utilidad de los modelos multifactoriales clásicos y abre la puerta a análisis futuros que incorporen más variables, otros datos o algoritmos avanzados orientados específicamente a series temporales.

Palabras clave:

Asset Pricing, Fama-French, Machine Learning, Random Forest, Mercado Europeo, Factores de Riesgo, Predicción Financiera.

1. INTRODUCCIÓN

Uno de los temas importantes en las finanzas modernas es entender los factores de rendimiento de los activos financieros. Desde mediados del siglo veinte, los académicos y los inversores han intentado responder una pregunta: ¿Cómo explicar los retornos de los activos? Aunque parece fácil contestar, la pregunta es mucho más compleja de lo que parece a simple vista.

La primera forma de dar una respuesta a esta pregunta fue el Capital Asset Pricing Model (CAPM) desarrollado por Sharpe y Lintner. Su propuesta se basaba en que los retornos esperados de un activo pueden explicarse únicamente por su exposición al riesgo sistemático del mercado gracias a la beta. Durante una década, este modelo dominó tanto la investigación académica como la gestión de inversiones pero con el paso del tiempo empezaron a aparecer pruebas que ponían en duda su capacidad de explicar las diferencias de los retornos entre activos.

A principios de los años noventa, Eugene Fama y Kenneth French realizaron una de las contribuciones más importantes en la historia de las finanzas. Demostraron que el CAPM no lograba explicar varios patrones persistentes, las anomalías de mercado como el efecto tamaño o el efecto valor. Para hacer frente a este problema, propusieron el modelo de tres factores incorporando dos nuevos componentes: SMB (Small Minus Big) y HML (High Minus Low). Este modelo supuso un avance decisivo, ya que lograba capturar variaciones en los retornos que el CAPM simplemente ignoraba.

Más de veinte años después, y como resultado de la acumulación de nuevas anomalías relacionadas con la rentabilidad y la inversión corporativa, Fama y French ampliaron su propuesta hasta conformar el modelo de cinco factores (2015). Este modelo incluye dos factores adicionales: la rentabilidad operativa (RMW) y el nivel de inversión (CMA). La literatura demuestra que esta versión más completa mejora la capacidad explicativa en la mayoría de mercados.

Sin embargo, los modelos lineales tradicionales tienen una limitación importante ya que asumen que la relación entre factores y retornos es lineal y estable en el tiempo. En un mundo financiero caracterizado por la complejidad, la velocidad y la abundancia de datos, esto podría ser demasiado restrictivo. Además, el desarrollo reciente de técnicas de Machine Learning ha despertado un gran interés en las finanzas cuantitativas. Modelos capaces de capturar relaciones no lineales, interacciones entre variables y patrones que

los métodos clásicos pasan por alto están transformando la forma en que se analiza la información financiera.

Gracias a esto surge la motivación de este Trabajo de Fin de Grado: evaluar si las técnicas de Machine Learning pueden mejorar la capacidad predictiva del modelo de cinco factores de Fama y French en el mercado europeo. Se compara el modelo lineal clásico que es estimado mediante regresión OLS con un modelo no lineal utilizado en predicción, el Random Forest Regressor.

Para ello, se construye un portafolio con siete acciones europeas de gran capitalización y se utilizan los factores mensuales oficiales de Fama-French para Europa. Después se entrenan ambos modelos utilizando datos desde el 2002 hasta el 2025 y posteriormente se evalúan fuera de muestra, asegurando un análisis realista y completo.

Los resultados, como se detallará más adelante, revelan un hallazgo interesante, que el modelo lineal de Fama-French supera al modelo de Machine Learning en precisión y ajuste, lo que sugiere que las relaciones entre factores y retornos siguen siendo fundamentalmente lineales. Esta conclusión va en línea con la literatura más reciente, que señala que los modelos de ML solo aportan valor sustancial cuando se dispone de un número muy grande de predictores o cuando existen no linealidades fuertes en los datos, algo que no parece darse en este caso.

Este trabajo se estructura de la siguiente manera. El Marco Teórico presenta los fundamentos conceptuales del asset pricing y la evolución de los modelos multifactoriales. La Revisión de la Literatura resume las aportaciones más relevantes en este campo, desde el modelo original de Fama-French hasta los avances recientes en Machine Learning. La Metodología describe los datos, los modelos estimados y los criterios de evaluación. A continuación, en Resultados se exponen los hallazgos y su interpretación económica. Y para terminar, las Conclusiones que recogen las principales aportaciones del estudio, sus limitaciones y sus futuras líneas de investigación.

2. MARCO TEÓRICO

El estudio del comportamiento de los precios de los activos financieros ha sido una de las cuestiones centrales en las finanzas modernas, por lo que entender por qué algunos activos ofrecen mayores rendimientos que otros y qué fuentes de riesgo son relevantes es fundamental tanto para la teoría financiera como para la inversión. Aunque la idea de que hay una relación entre riesgo y retorno parece lo normal, entender esta relación ha necesitado muchos años de investigación.

En este capítulo se ve como han ido evolucionando los modelos de valoración de activos desde los planteamientos más simples hasta los mas complejos que han surgido en los últimos años. El objetivo es construir un marco conceptual para poder entender por qué el modelo de Fama y French ha sido tan importante, cuáles son sus límites y por qué aporta valor compararlo con un modelo de Machine Learning como el Random Forest.

2.1. Origen y principios del Asset Pricing

El asset pricing nace ya que los inversores deberían de ser recompensados por el riesgo que asumen. Pero la clave está en determinar qué se entiende por riesgo y, sobre todo, qué parte de ese riesgo es relevante para los precios de los activos. A lo largo del tiempo, la mayoría de estudios están de acuerdo en que únicamente el riesgo sistemático, que es el que no puede eliminarse mediante diversificación, tiene que estar conectado a una prima de riesgo. Esta idea separa el ruido propio de cada empresa del componente común que afecta a todos.

La teoría financiera moderna ha intentado descubrir qué factores explican este riesgo y cómo se relaciona cada activo con ellos. De ahí han surgido varios modelos, algunos más teóricos y otros más prácticos, que buscan traducir este concepto en ecuaciones aplicables a los datos reales.

2.2. El CAPM y sus limitaciones

El primer intento relacionr los factores con el riego fue el Capital Asset Pricing Model (CAPM). Su gran aportación fue identificar el riesgo sistemático gracias a único parámetro, la beta del mercado. Según este modelo, un activo debería tener un rendimiento esperado mayor cuando su beta es superior, es decir, si se movía de forma

más intensa con el mercado. La simplicidad del CAPM lo convirtió rápidamente en el estándar tanto para analizar carteras como para tomar decisiones corporativas.

Sin embargo, la simpleza del CAPM supone un problema, que es incapaz de explicar ciertos patrones que se repiten. Muchos estudios publicados desde los años setenta han demostrado que empresas pequeñas o que crecían poco en términos contables generaban retornos anómalamente altos que no encajaban con la teoría. Estos patrones empezaron a conocerse como anomalías de mercado y representaron una de las primeras evidencias de que la relación entre el riesgo y el retorno es más compleja de lo que sugiere el CAPM.

Además, el modelo se apoyaba en ciertos supuestos como la existencia de un mercado perfectamente eficiente o que las expectativas son iguales, que no se suelen cumplir en la realidad. Esto no invalida su valor teórico, pero sí limita su utilidad práctica abriendo la puerta a modelos más complejos.

2.3. El modelo de tres factores de Fama y French (1993)

A principios de los años noventa, Eugene Fama y Kenneth French realizaron un avance que redefinió por completo la forma de entender el riesgo del mercado. En su modelo de tres factores añadieron dos componentes adicionales al factor de mercado, un factor relacionado con el tamaño de las empresas y otro con el ratio valor contable/valor de mercado. Asi consiguieron explicar muchos de los problemas que tenía el CAPM.

La lógica detrás de estos factores era que las empresas pequeñas y las empresas con altos ratios valor contable/valor de mercado se comportaban de manera diferente al resto precisamente porque estaban sujetas a riesgos específicos que el CAPM no tenia en cuenta. El modelo de tres factores no solo tuvo un gran impacto, sino que supuso una reinterpretación profunda del riesgo. Si el CAPM sugería que todo podía explicarse con el mercado, Fama y French demostraron que el mercado era solo una parte del conjunto de riesgos comunes.

Aun con esta mejora quedaban todavía varios aspectos que explicar. Otros patrones relacionados con la inversión corporativa o la rentabilidad operativa seguían sin tener una explicación teórica.

2.4. El modelo de cinco factores de Fama y French (2015)

Más de veinte años después, Fama y French ampliaron su propuesta con un modelo de cinco factores en el que añadían dos nuevas variables, una relacionada con la rentabilidad operativa de las empresas (RMW) y otra con sus políticas de inversión (CMA). Este cambio se apoyaba en teorías corporativas que conectan los beneficios y las decisiones de inversión con los valores fundamentales de las empresas.

La ampliación del modelo reflejaba un cambio en el enfoque, ya no se trataba de añadir factores solo porque funcionaran en la práctica, sino de conectar los retornos con conceptos económicos reales. Las empresas más rentables tienden a ser mejores inversiones porque generan flujos de caja más estables y las empresas que invierten de manera muy agresiva pueden acabar perdiendo valor si se acaban equivocando. Estas consideraciones económicas daban un sentido profundo a los factores.

Sin embargo, con el modelo de cinco factores también se generó un debate interesante. En muchos datasets, el factor HML que era una de las bases en el modelo de tres factores, perdía relevancia cuando se añadían RMW y CMA. Esto ponía en duda la persistencia del efecto valor y abría la pregunta de que si algunas anomalías podrían haber sido resultado de métodos estadísticos poco robustos o de sesgos en los datos.

2.5. Anomalías, eficiencia y multifactorialidad

Los modelos de Fama y French han sido aclamados por su capacidad para capturar muchas anomalías de mercado, pero la literatura reciente sugiere que parte de estas anomalías pueden deberse a problemas de replicación. En un estudio de Hou, Xue y Zhang del 2017 con más de 400 señales, se concluye que una gran parte de las anomalías desaparecen cuando se aplican metodologías más duras, como el control por microcaps o el uso de carteras ponderadas por valor.

Esto ha generado dos posiciones distintas. Por un lado, los que dicen que el que haya tantas anomalías refleja que el mercado es complejo y que se necesitan modelos cada vez más multifactoriales. Por otro lado, están los que apoyan metodologías más sólidas diciendo que los mercados son más eficientes de lo que se pensaba inicialmente.

De todas formas, los modelos multifactoriales tradicionales comparten una estructura lineal que limita su capacidad para identificar las relaciones más complejas entre variables.

2.6. Limitaciones inherentes a los modelos lineales

A pesar del éxito que han tenido los modelos de Fama y French, se siguen viendo varias limitaciones. La primera es que suponen que la relación entre cada factor y los retornos es lineal. Esto es útil desde el punto de vista econométrico, pero probablemente simplifica demasiado la realidad ya que los mercados financieros están llenos de comportamientos no lineales y cambios que no pueden representarse mediante ecuaciones lineales.

Una segunda limitación es que los modelos lineales asumen que los coeficientes son constantes en el tiempo. Esto no suele pasar en los mercados reales ya que la sensibilidad de una empresa a los distintos factores puede variar dependiendo del ciclo económico, de la volatilidad o de su propia evolución.

Finalmente, los modelos lineales suelen incluir pocos factores y la literatura reciente muestra que existen cientos de características capaces de predecir retornos en distintos contextos, desde datos contables hasta información extraída de noticias o informes financieros. Integrar toda esta información con un enfoque lineal es casi imposible.

Estas limitaciones han motivado la búsqueda de métodos nuevos, sobre todo en el ámbito del Machine Learning.

2.7. La aportación del Machine Learning al Asset Pricing

El Machine Learning ha introducido una nueva forma de aproximarse al análisis financiero. A diferencia de los modelos lineales tradicionales, los algoritmos de ML en lugar de asumir cómo deberían relacionarse los factores con los retornos, dejan que los propios datos revelen patrones.

Modelos como Random Forest, Gradient Boosting o las redes neuronales permiten capturar interacciones complejas, dependencias no lineales y relaciones de alta

dimensión. Este enfoque resulta especialmente útil cuando se trabaja con cantidades muy grandes de datos o cuando se sospecha que el comportamiento de los activos no sigue una estructura lineal simple.

Así, mientras que modelos como el de Fama y French son interpretables y más rígidos, los modelos de Machine Learning priorizan la capacidad predictiva y la flexibilidad. El tema actual en la literatura no gira tanto en torno a cuál es mejor, sino a cuándo conviene utilizar cada uno y qué aportan realmente en contextos diferentes.

Comparar ambos enfoques en un entorno practico no permite evaluar hasta qué punto los modelos no lineales captan información adicional relevante o si, por el contrario, la estructura lineal propuesta por Fama y French sigue siendo suficiente para describir los rendimientos de los activos.

3. REVISIÓN DE LA LITERATURA

La investigación en asset pricing ha tenido una gran evolución durante las últimas décadas. Desde los primeros modelos basados en un único factor hasta los modelos actuales que integran cientos de variables y emplean técnicas avanzadas de Machine Learning, el asset pricing ha experimentado una gran transformación metodológica y conceptual. Este capítulo da una revisión de la literatura más relevante, con especial atención a los trabajos en los que se basa este estudio de los que hablaremos a continuación.

3.1. Common Risk Factors in the Returns on Stocks and Bonds (1993)

El trabajo de Fama y French (1993) marcó un antes y un después en la literatura financiera. Hasta ese momento, el CAPM seguía siendo el modelo dominante a pesar de que cada vez se notaban más sus límites. Lo que hicieron Fama y French fue algo más que añadir factores adicionales al modelo clásico, propusieron una nueva forma de comprender las primas de riesgo basándose en evidencias sólidas y en patrones que se repetían en los datos.

En este estudio, se vio que los retornos de las acciones y de los bonos podían explicarse de manera más robusta si además del riesgo de mercado se incluían dos factores adicionales, uno relacionado con el tamaño de la empresa y otro con la relación valor contable/valor de mercado. Esta contribución no solo explicaba anomalías históricas como el efecto tamaño o valor, sino que proponía que estas anomalías representaban riesgos recurrentes que los inversores realmente enfrentaban.

Con ello, Fama y French no solo ampliaron el modelo tradicional, sino que también ofrecieron un marco más realista que integraba características observables de las empresas. Gracias a su impacto, se convirtió en una referencia y dio inicio a los modelos multifactoriales.

3.2. A Five-Factor Asset Pricing Model (2015)

Unos veinte años más tarde, Fama y French publicaron uno de los artículos más influyentes de la última década donde revisaban su anterior propuesta, el modelo de cinco factores. En este trabajo, los autores reconocen que su modelo de tres factores no

capturaba completamente la relación entre los rendimientos medios y ciertos aspectos fundamentales de las empresas. En concreto, observaban que la rentabilidad operativa y la política de inversión jugaban un papel muy importante en la valoración de los activos.

La inclusión de los factores RMW y CMA servía para dos cosas. Por un lado, para aportar explicaciones más unidas a la teoría económica de la empresa. Y por otro lado, para darle sentido a anomalías que el modelo de tres factores no podía explicar. Los autores muestran que el nuevo modelo reduce considerablemente los errores de pricing.

Sin embargo, su propuesta también planteó un debate importante. El factor HML, considerado esencial en el modelo original, pasaba a tener un papel marginal en presencia de los nuevos factores. Esto generó discusiones en torno a si el efecto valor había perdido relevancia en los mercados modernos o si el modelo de cinco factores estaba absorbiendo información de forma redundante.

Este cambio es uno de los motivos por los que resulta interesante evaluar la capacidad explicativa de los cinco factores en mercados distintos del estadounidense, como es el caso europeo en este TFG.

3.3. Comparing Cross-Section and Time-Series Factor Models (2019)

La literatura sobre asset pricing se divide tradicionalmente en dos enfoques. El de los modelos transversales(cross-section) que explican las diferencias en los retornos medios entre activos. Y el de los modelos temporales(time-series) que explican la evolución en el tiempo del retorno de un activo en función de su exposición a los factores.

Este estudio analiza la relación entre ambos enfoques. Una de sus conclusiones principales es que no siempre utilizan la misma información, ni tienen por qué llevar a los mismos resultados. Esto es relevante para este TFG porque el análisis que se realiza es de tipo time-series como el método de Fama-French.

Los autores señalan que los modelos temporales suelen ofrecer mejores ajustes cuando se evalúan carteras agregadas, mientras que los modelos transversales son más exigentes cuando se intentan justificar las diferencias entre activos individuales. Esta distinción implica que un modelo que funciona bien en series temporales, como el que se emplea aquí, no tiene por qué explicar igual la dispersión transversal de retornos.

3.4. Replicating Anomalies (2017)

El estudio de Hou, Xue y Zhang es uno de los análisis más completos y exigentes publicados en los últimos años. En este estudio se recopilaron 447 anomalías documentadas en estudios anteriores y las replicaron utilizando metodologías consistentes y corregidas. Sus conclusiones fueron que más del 60% de las anomalías no se replicaban de forma significativa, y el porcentaje aumentaba al 85% cuando se aplicaban estándares estadísticos más estrictos.

El estudio pone mucha importancia en el problema del p-hacking, señalando que muchos resultados publicados podrían deberse a decisiones metodológicas ad hoc o a selecciones de muestras favorables. Esto no solo cuestiona la validez de muchas anomalías, sino que también genera dudas sobre cuántos de los factores propuestos en décadas anteriores representan verdaderos riesgos y cuantos han sido amañados.

Además, los autores muestran que el modelo q-factor explica un número elevado de anomalías que se repiten. Este resultado refuerza la importancia de los factores relacionados con RMW y CMA, los mismos que forman parte del modelo de cinco factores de Fama y French.

El mensaje que quiere transmitir este estudio es que no todas las señales son válidas, y que los modelos de valoración deben apoyarse en fundamentos económicos más sólidos y en datos más rigurosos. Este argumento cobra importancia en un contexto donde el Machine Learning tiende a generar gran cantidad de predictores sin tener por qué garantizar su importancia teórica.

3.5. Unraveling Asset Pricing with AI (2023)

En los últimos años, el asset pricing ha vivido una nueva transformación gracias al avance de la inteligencia artificial. Este estudio ofrece una revisión de más de 780 estudios que aplican técnicas de Machine Learning y Deep Learning a la predicción de retornos. Según los autores, la incorporación de algoritmos como Random Forest, redes neuronales o modelos de grafos ha permitido identificar patrones mucho más complejos que los modelos tradicionales.

Sin embargo, también dicen que el rendimiento de estas técnicas depende de mucho de la calidad de los datos y de la capacidad del modelo para capturar estructuras no lineales reales. En particular, destacan que los factores temporales de los mercados financieros contienen un mucho ruido y que, si no se gestiona adecuadamente, los modelos avanzados pueden sobreajustarse fácilmente. Esto es coherente con la evidencia de este TFG, donde el modelo Random Forest no supera al modelo lineal.

Otro punto importante es la creciente diferencia entre modelos predictivos y modelos explicativos. Mientras que los modelos lineales permiten interpretar los coeficientes y entender la lógica económica, los modelos complejos pueden resultar opacos para académicos y reguladores. En este sentido, el estudio recomienda combinar modelos lineales para la explicación y modelos machine learning para la exploración y predicción.

El artículo destaca finalmente la necesidad de incorporar principios económicos en el diseño de los algoritmos, ya que un modelo puramente estadístico sin restricciones económicas corre el riesgo de identificar patrones falos.

3.6. Síntesis de la literatura y aportación de este estudio

La literatura revisada refleja un debate continuo sobre cómo deben modelizarse los retornos de los activos. Por un lado, los modelos multifactoriales de Fama y French han demostrado ser robustos y sorprendentemente precisos. Por otro lado, lo estudios recientes muestran que muchos factores y anomalías no resisten una replicación rigurosa, lo que refuerza la necesidad de enfoques teóricamente fundamentados.

Al mismo tiempo, la aparición del Machine Learning ha abierto nuevas posibilidades, pero también nuevos desafíos. Se ha demostrado la capacidad de estos algoritmos para capturar no linealidades y relaciones complejas, pero su desempeño depende en gran medida del contexto y la estructura de los datos.

En este TFG, se enfrentan el modelo clásico de cinco factores de Fama-French y el Random Forest, representando la tradición lineal frente a la flexibilidad del Machine Learning. La literatura sugiere que los modelos lineales pueden ser sorprendentemente difíciles de superar cuando los factores están bien construidos y tienen un fundamento

económico sólido. Los resultados de este estudio, como veremos más adelante, ofrecen una confirmación de esta idea.

4. METODOLOGÍA

Este capítulo describe el proceso seguido para llevar a cabo el análisis practico. El objetivo es proporcionar una visión clara y transparente de cómo se han obtenido los datos, cómo se han transformado, qué modelos se han estimado y con qué criterios se han evaluado. La metodología en un estudio de este tipo es tan importante como los resultados, ya que determina la validez y la replicabilidad del análisis.

La aproximación seguida combina rigor académico usando los factores originales de Fama French y estimar un modelo lineal estándar con elementos más modernos alineados con la literatura de predicción financiera, como la separación entre datos de entrenamiento y prueba y la implementación de un modelo de Machine Learning no lineal. Esto permite contrastar ambos enfoques bajo un procedimiento experimental.

4.1. Datos: fuentes, periodo y justificación de la muestra

La base de datos utilizada en este estudio proviene de dos fuentes principales. Por un lado, los factores de Fama-French se descargaron del repositorio oficial mantenido por Kenneth French en Dartmouth College. Se empleó la versión Europe 5 Factors, que contiene datos mensuales del mercado europeo, incluyendo el factor de mercado (Mkt-RF), el tamaño (SMB), el valor (HML), la rentabilidad operativa (RMW), la inversión (CMA) y la tasa libre de riesgo (RF). La ventaja de utilizar estos factores es que siguen la metodología estándar utilizada en cientos de estudios académicos, lo que garantiza comparabilidad y rigor.

Por otro lado, los precios de las acciones se obtuvieron mediante la librería yfinance, que permite descargar datos de mercados internacionales directamente de la red. Para formar el portafolio se seleccionaron siete empresas europeas de gran capitalización pertenecientes a índices como el IBEX 35, CAC 40 o DAX. Entre los valores elegidos se encuentran Santander, BBVA, Inditex, Iberdrola, Telefónica, Repsol y ACS. Esta selección responde a varios criterios como liquidez suficiente, continuidad de datos, pertenencia a mercados desarrollados y representatividad sectorial dentro de Europa.

El periodo analizado abarca desde febrero de 2002 hasta octubre de 2025, lo que proporciona un total de 284 observaciones mensuales. Esta longitud temporal es adecuada

para estimar modelos de series temporales en asset pricing, ya que proporciona suficiente variabilidad en los factores y permite evaluar la estabilidad de los coeficientes.

Es importante que en finanzas los análisis suelen realizarse con datos mensuales y no diarios. Esto se debe a que los factores se construyen mensualmente y a que los modelos de Fama-French buscan explicar rendimientos medios y no cambios diarios que están más influenciados por el ruido de mercado.

4.2. Construcción del portafolio y transformación de los precios

Una vez descargados los datos de precios, se procedió a transformarlos en retornos mensuales. Como los precios originales son diarios, se ha seleccionado el precio de cierre correspondiente al último día de cada mes. Esta es la técnica estándar para convertir frecuencia diaria en mensual, cambiando lo menos posible la estructura temporal de la serie.

Los retornos se calcularon mediante logaritmos (log-returns), una práctica común en econometría financiera porque facilitan la interpretación, tienden a tener mejor comportamiento estadístico (especialmente cuando se combinan múltiples activos) y se van sumando en el tiempo. Cada acción genera así una serie temporal de retornos mensuales, que luego se promedian para obtener el retorno del portafolio.

La decisión de hacer un portafolio ponderado es por varios motivos. En primer lugar, permite evitar que dominen las empresas más grandes, algo que podría distorsionar los resultados y reducir la representatividad de los factores. Y en segundo lugar, los portafolios ponderados son muy comunes en la literatura académica, especialmente en estudios que buscan comparar modelos en condiciones controladas.

4.3. Integración de los factores y el portafolio: limpieza y alineación temporal

Una parte muy importante de la metodología es alinear correctamente los retornos del portafolio con los factores mensuales. Pequeños errores en este paso pueden hacer surgir problemas de sincronización que afectan a los resultados. Para evitarlo, se han ajustado las fechas para que tanto las series de factores como el retorno del portafolio estuvieran definidas exactamente en los mismos meses.

Una vez combinadas las series temporales, cualquier observación con valores faltantes fue eliminada. Dado que el periodo común entre las acciones y los factores es muy grande, el impacto de este filtrado fue reducido. El resultado final es un dataset completamente limpio y adecuado para hacer el estudio.

4.4. Especificación y estimación del modelo OLS

El modelo lineal estimado sigue la ecuación del modelo de cinco factores de Fama y French:

$$R_{p,t} - R_{f,t} = \alpha + \beta(Mkt - RF)_t + sSMB_t + hHML_t + rRMW_t + cCMA_t + \varepsilon_t$$

donde $R_{p,t}$ representa el retorno del portafolio en el mes t, y $R_{f,t}$ es la tasa libre de riesgo. La ecuación se estima mediante mínimos cuadrados ordinarios (OLS), utilizando la librería statsmodels.

En este contexto, la interpretación de los coeficientes es directa, cada parámetro indica cómo cambia el retorno del portafolio por una unidad de variación en el factor correspondiente. El término α es especialmente importante, ya que indica si existe algún retorno no explicado por los factore, lo que significaría que hay anomalías o señales adicionales.

Para evaluar el rendimiento del modelo se calculan dos datos. El R^2 indica qué porcentaje de la variación de los retornos del portafolio puede explicarse mediante los factores. Y el MSE (Mean Squared Error) que mide el error medio cuadrático en el conjunto de prueba, ofreciendo una visión más centrada en la calidad predictiva.

4.5. El modelo Random Forest: estructura, lógica y configuración

El segundo modelo estimado es un Random Forest Regressor, una técnica de Machine Learning basada en árboles de decisión. A diferencia del modelo lineal, el Random Forest no impone una estructura funcional previa sino que construye múltiples árboles que aprenden patrones de los datos y luego promedia sus predicciones.

Las razones para usar un Random Forest en lugar de otros modelos más complejos son varias. En primer lugar, es un algoritmo robusto que maneja bien relaciones no lineales y posibles interacciones entre factores. En segundo lugar, requiere menos datos

para entrenarse que modelos como redes neuronales, lo cual es importante en un estudio con una matriz de predictores pequeña. Finalmente, ofrece medidas de importancia de variables, lo que facilita la interpretación relativa de los factores.

Los hiperparámetros usados como el número de árboles, la profundidad máxima y el tamaño mínimo de hoja se eligieron para tener un buen equilibrio entre la flexibilidad y el riesgo de sobreajuste. El algoritmo se entrenó únicamente con los datos del conjunto de entrenamiento para evaluar su capacidad predictiva real.

4.6. Separación entre entrenamiento y prueba

Una de las diferencias entre la econometría tradicional y el Machine Learning es la importancia de la validación fuera de la muestra. Para eso, se dividió el dataset en una parte de entrenamiento (80%) y otra de prueba (20%). Esta división mantiene el orden temporal de las observaciones, lo que evita que la información del futuro se vaya hacia el pasado, lo cual invalidaría cualquier conclusión.

El conjunto de entrenamiento se utilizó para estimar los coeficientes del modelo lineal y para que el Random Forest se aprendiera los patrones de los datos. El conjunto de prueba, que está formado por los últimos 57 meses, permite evaluar el rendimiento real de ambos modelos en condiciones comparables.

4.7. Métricas de evaluación y criterios comparativos

Para comparar ambos modelos se han usado dos métricas principales, el R^2 y el MSE. Aunque el R^2 es una medida que va tradicionalmente con ajuste del modelo lineal, también se puede usar como indicador de la capacidad explicativa del Random Forest. Por su parte, el MSE ofrece una forma de evaluar hasta qué punto las predicciones del modelo se acercan a los retornos observados.

Además, se ha analizado la importancia de los factores en el Random Forest lo que deja evaluar si el modelo no lineal identifica una jerarquía similar a la del modelo lineal o si detecta patrones distintos, lo cual podría dar relaciones no lineales entre factores y retornos.

La última parte de la metodología consiste en comparar los gráficos de predicciones frente a retornos reales, así como los coeficientes o importancias estimadas. Este análisis visual complementa las métricas cuantitativas y ayuda a ver si alguno de los modelos tiene problemas de estabilidad o desviaciones sistemáticas.

4.8. Implementación práctica en Python

Todo el análisis se implementó en Python utilizando librerías especializadas como pandas, numpy, statsmodels, sklearn y matplotlib. El código está completamente documentado y se incluye en los anexos del TFG. La ventaja de Python sobre otros programas como R o MATLAB es su versatilidad para combinar modelos econométricos clásicos con algoritmos de Machine Learning dentro del mismo framework, lo que hace mas facil una comparación directa y rigurosa.

5. RESULTADOS

El objetivo de este capítulo es presentar, analizar e interpretar los resultados obtenidos de la estimación del modelo de cinco factores de Fama-French mediante regresión lineal y del modelo no lineal Random Forest. La comparación entre los dos métodos permite ver desde una perspectiva práctica si la complejidad adicional del Machine Learning da más valor para explicar los retornos de un portafolio europeo frente al marco lineal tradicional.

Los resultados se describen de forma clara y detallada, integrando no solo métricas cuantitativas, sino que también interpretaciones económicas que ayudan a comprender el comportamiento de los factores y la estructura de los datos. Este capítulo es la pieza central del análisis, ya que permite extraer conclusiones sobre la relevancia práctica de los modelos y sobre la validez de sus supuestos.

5.1. Estadísticos descriptivos y comportamiento general de los datos

Antes de hablar de los modelos, hay que contextualizar la naturaleza de los datos usados. El portafolio ponderado formado por siete acciones europeas tiene un comportamiento moderado, ni es demasiado volátil ni es demasiado estable. Esto se debe a que está basado en empresas de gran capitalización, relativamente defensivas y pertenecientes a distintos sectores (financiero, energía, construcción, telecomunicaciones, textil y utilities).

Durante el periodo del 2002 al 2025, el portafolio recoge varios eventos como la crisis financiera de 2008, la crisis de deuda europea, la pandemia de 2020, los efectos del repunte inflacionario en 2022–2023, y su fase de recuperación.

Estos sucesos generan una variabilidad suficiente como para que los factores de Fama-French capten distintos regímenes de mercado, lo cual es relevante para obtener conclusiones sólidas.

También los factores presentan una dinámica típica:

- El factor de mercado muestra los movimientos amplios y persistentes esperados.
- SMB y HML reflejan ciclos conocidos: el tamaño y el valor han tenido periodos de fuerte presencia y otros de debilidad.

- RMW y CMA muestran patrones asociados a rentabilidad e inversión, alineados con la teoría económica.

En conjunto, los datos presentan características adecuadas para un análisis de asset pricing que son la variabilidad suficiente, la ausencia de valores extremos no plausibles y una estructura temporal limpia y consistente.

5.2. Resultados del modelo OLS: el rendimiento del enfoque lineal

La estimación del modelo de cinco factores mediante OLS ofrece un ajuste sólido y consistente con la literatura. La métrica clave, el R^2 obtenido en el conjunto de prueba, tiene un valor de 0.6716, lo que indica que el modelo es capaz de explicar aproximadamente el 67% de la variación temporal en los retornos del portafolio europeo. Se trata de un valor elevado para un modelo lineal en asset pricing, especialmente considerando la simplicidad del portafolio empleado.

El MSE del modelo alcanza un valor de 0.000597, relativamente bajo, lo que confirma que las predicciones de la regresión no se desvían excesivamente de los retornos reales. Esto es importante ya que, aunque el asset pricing tradicional se centra más en explicar la media de los retornos que en predecir series temporales, la estabilidad del ajuste refuerza la validez del modelo.

Uno de los resultados más relevantes es que el término constante (alpha) es prácticamente cero. Esto indica que el modelo no deja un residuo sistemático sin explicar. En otras palabras, no parece haber anomalías adicionales significativas que no estén capturadas por los factores incluidos en la regresión. Esta ausencia de alpha refuerza la idea, defendida por Fama y French, de que los cinco factores incorporan la mayoría de las fuentes relevantes de riesgo sistemático.

Además, los coeficientes estimados siguen la intuición económica:

- El factor de mercado presenta el coeficiente más elevado con diferencia.
- SMB y HML muestran contribuciones más moderadas.
- RMW y CMA presentan coeficientes coherentes con la teoría (rentabilidad e inversión importan, pero no dominan).

En conjunto, el modelo OLS ofrece una explicación sólida, estable e interpretable del comportamiento del portafolio europeo.

5.3. Resultados del modelo Random Forest: la alternativa no lineal

El modelo Random Forest presenta un rendimiento inferior al del modelo lineal tanto en términos de R^2 como de MSE. El R^2 se sitúa en 0.6095, lo que supone una caída de algo más de seis puntos porcentuales respecto al modelo OLS. Aunque el ajuste sigue siendo razonable para un modelo no lineal con pocos predictores, confirma que en este contexto la estructura lineal de Fama-French parece capturar de manera más eficaz la relación entre los factores y los retornos.

El MSE del Random Forest asciende a 0.000710, un poco más alto que el del modelo lineal. A pesar de que la diferencia pueda parecer pequeña, en el contexto de un portafolio con retornos que rara vez superan unos pocos puntos porcentuales mensuales esta diferencia es significativa.

Uno de los elementos más interesantes del Random Forest es su medida de importancia relativa de los factores. Según el modelo, el factor Mkt-RF es el más relevante, acumulando casi la mitad de la importancia total. El segundo factor en importancia es HML, seguido por RMW, SMB y, finalmente, CMA. Este orden es importante porque dice que los factores tradicionales mantienen un peso que no cambia demasiado de la intuición dada por el modelo lineal.

La importancia tan elevada del factor de mercado es llamativa. En cierto modo, sugiere que la mayor parte de la dinámica del portafolio está unida a movimientos globales del mercado europeo, lo que tiene sentido con la selección de empresas grandes y diversificadas.

5.4. Comparación directa entre OLS y Random Forest

La comparación directa entre ambos modelos permite evaluar la utilidad relativa de cada enfoque. Los resultados muestran que el modelo OLS tiene una capacidad explicativa superior y menor error de predicción. Esto significa que incorporar no linealidades y permitir interacciones complejas entre los factores no mejora la calidad del modelo.

La diferencia en R^2 entre ambos modelos es de 0.0621, lo que es significativo en un estudio de asset pricing. Aunque ambos modelos superan bastante el 0.50, que suele considerarse un umbral razonable en predicción financiera mensual, el modelo lineal tiene una ventaja clara. Además, la diferencia en el MSE refuerza esta idea, confirmando que el Random Forest no capta patrones adicionales que el modelo OLS esté dejando fuera.

En consecuencia, la comparación sugiere que, al menos para este portafolio y este período temporal, el modelo de Fama-French es capaz de ver la estructura de los retornos sin necesidad de recurrir a enfoques más complejos.

5.5. Interpretación económica de los resultados

Más allá de las métricas numéricas, los resultados permiten obtener una lectura más profunda sobre la naturaleza de los retornos de los activos europeos. En primer lugar, la superioridad del modelo lineal dice que las relaciones entre los factores y los retornos son esencialmente lineales. Esto no implica que no existan fenómenos no lineales en los mercados, pero sí que indica que en un portafolio ponderado y com cinco factores principales, las no linealidades no parecen ser determinantes.

En segundo lugar, la ausencia de alpha significativo confirma la eficiencia relativa del mercado europeo en su conjunto. Si el modelo lineal fuera claramente insuficiente, podríamos decir que el modelo no lineal captura alguna estructura adicional. Sin embargo, parece razonable concluir que los factores incluidos en el modelo de Fama-French contienen la mayor parte de la información relevante.

Por último, la importancia tan elevada del factor de mercado, tanto en el modelo lineal como en el Random Forest, sugiere que las empresas del portafolio comparten una exposición común muy fuerte al ciclo económico europeo. Esto tiene sentido ya que las compañías seleccionadas son líderes en sus respectivos sectores y están fuertemente integradas en la economía europea.

5.6. Análisis visual: coherencia entre predicciones y retornos observados

Los gráficos de predicciones frente a retornos reales permiten corroborar visualmente las conclusiones numéricas. Tanto en el modelo OLS como en el Random Forest, las líneas de predicción siguen razonablemente bien los movimientos generales del portafolio.

No obstante, se ve que el Random Forest tiende a suavizar demasiado las variaciones y responde peor a cambios bruscos en los retornos, especialmente cuando hay volatilidad. En contraste, el modelo OLS muestra una capacidad un poco mejor para adaptarse a los movimientos mensuales, lo que probablemente refleja la influencia del factor de mercado.

En los gráficos de dispersión, se ve claramente cómo las predicciones del modelo OLS se alinean más con la línea de 45°, mientras que el Random Forest tiene una mayor dispersión. Esto confirma que las predicciones lineales son más precisas en términos absolutos.

5.7. Síntesis de los hallazgos del capítulo

Los resultados de este capítulo tienen una conclusión clara, en el contexto analizado el modelo de Fama-French de cinco factores supera al modelo Random Forest tanto en capacidad explicativa como en precisión predictiva. A pesar de su mayor complejidad, el modelo de Machine Learning no ofrece una mejora y resulta menos eficiente.

Esto sugiere que la estructura de los retornos en el portafolio europeo sigue patrones bastante lineales y que los factores tradicionales dan información suficiente para explicar su comportamiento. También refuerza la literatura reciente que explica que los modelos de ML superan a los modelos lineales solo cuando se dispone de una gran cantidad de predictores o cuando la estructura de los datos presenta no linealidades significativas.

6. DISCUSIÓN

Los resultados que se ven en el capítulo anterior dejan hacer una reflexión sobre la naturaleza de los modelos de valoración de activos, la utilidad real de los factores de Fama-French, y del papel que tiene el Machine Learning en el ámbito financiero. Este capítulo sirve para poner en contexto los hallazgos, discutir sus implicaciones teóricas y prácticas, y evaluar hasta donde los resultados obtenidos encajan con la literatura.

El análisis llevado a cabo en este trabajo muestra que el modelo clásico lineal de cinco factores no solo funciona bien, sino que supera en capacidad predictiva al modelo Random Forest. Dada la creciente popularidad de los modelos no lineales en asset pricing, este resultado merece una discusión detallada.

6.1. ¿Por qué el OLS supera al Random Forest en este estudio?

Cuando se comparan modelos de Machine Learning con modelos lineales tradicionales, se suele esperar que los primeros capten relaciones más complejas y que den predicciones mejores. Sin embargo, esto no ha ocurrido en este trabajo. Existen varias razones que explican por qué el modelo lineal de Fama-French presenta un rendimiento superior.

La primera razón tiene que ver con la estructura del dataset. Los factores de Fama-French están construidos precisamente para explicar retornos medios y relaciones sistemáticas lineales. No son variables arbitrarias, sino el resultado de décadas de refinamiento empírico y teórico. Si estos factores ya representan la verdadera estructura de riesgo, no es sorprendente que un modelo lineal sea capaz de capturar la relación prácticamente en su totalidad.

En segundo lugar, el Random Forest tiene dificultades cuando el número de predictores es pequeño. Este tipo de algoritmo sirve sobre todo en contextos con decenas o cientos de variables, donde puede explorar posibles interacciones. En este análisis, solo existen cinco factores explicativos, lo que reduce mucho su ventaja.

Un tercer elemento clave es que el portafolio analizado es relativamente estable y diversificado. Como no hay estructuras altamente no lineales es poco probable que el Machine Learning mejore el rendimiento de un modelo lineal. Los árboles, al ser más

sensibles al ruido, pueden incluso suavizar demasiado las predicciones o sobrerreaccionar cuando hay valores atípicos.

Por último, un aspecto importante es que el Random Forest no está diseñado para capturar bien estructuras de series temporales sin modificaciones adicionales. Aunque puede aplicarse a datos temporales, no tiene memoria histórica ni incorpora dinámicas temporales de forma explícita. Por otro lado, el modelo OLS asume una estructura estática pero lineal, lo cual parece más consistente con la naturaleza de los datos en este caso.

6.2. Interpretación de los resultados en el contexto del mercado europeo

Los resultados también permiten reflexionar sobre el comportamiento del mercado europeo durante el periodo analizado. La predominancia del factor de mercado en ambos modelos sugiere que las grandes empresas europeas están fuertemente alineadas con el ciclo financiero. El portafolio seleccionado, que incluye compañías líderes en sectores estratégicos, capta muy bien esa tendencia.

El papel de los factores HML, RMW y CMA también resulta ilustrativo. Aunque la literatura reciente discute la pérdida de relevancia del factor valor cuando se introducen los factores de rentabilidad e inversión, el análisis de este TFG muestra que HML aún desempeña un papel significativo en Europa. Esto se puede deber a que el mercado europeo mantiene una estructura sectorial diferente a la estadounidense, donde la tecnología ha dominado las últimas dos décadas.

La importancia moderada de RMW y CMA coincide con lo que se ve en la literatura europea, que la rentabilidad operativa influye en los retornos, pero su impacto no es tan importante como en Estados Unidos, donde el factor profitability fue uno de los principales descubrimientos de Fama y French en su estudio de 2015.

En conjunto, los resultados sugieren que el modelo de cinco factores es suficientemente estable y robusto para explicar retornos en el mercado europeo, aunque con la diferencia de que el factor valor mantiene aquí un peso considerable.

6.3. Conexión con la literatura académica

Uno de los aspectos más interesantes de este estudio es que los resultados están muy alineados con lo que se dice en la literatura más reciente. En particular, el hecho de que el modelo OLS supere al Random Forest coincide con varios estudios que dicen que los modelos lineales funcionan muy bien en entornos con pocos predictores y datos relativamente limpios.

La literatura de Hou, Xue y Zhang (2017) mostraba que muchas anomalías desaparecen cuando se trabajan con datasets depurados, lo que refuerza la idea de que los factores que realmente importan son pocos y estables. Por su parte, estudios como los revisados en *Unraveling Asset Pricing with AI* señalan que el ML ofrece mejoras sustanciales únicamente cuando se trabaja con bases de datos amplias, con muchas características y con un volumen de información mucho mayor al utilizado en este estudio.

Esto sugiere que, al menos en el contexto de factores tradicionales de Fama-French, la sofisticación metodológica del ML no aporta una mejora clara en términos predictivos. Dicho de otro modo: cuando el modelo teórico está bien especificado, la complejidad adicional no implica necesariamente una ganancia en poder predictivo.

Los resultados también coinciden con otro punto importante de la literatura, la interpretabilidad. Los modelos lineales permiten entender de manera clara y directa cómo cada factor contribuye al retorno del portafolio. En cambio, el Random Forest, aunque ofrece medidas de importancia, no permite reconstruir relaciones exactas ni entender por completo la estructura interna del modelo.

Este aspecto tiene especial importancia en un entorno donde reguladores e inversores demandan modelos transparentes, sobre todo en áreas sensibles como la asignación de capital o la gestión del riesgo.

6.4. Implicaciones para la teoría y para la práctica financiera

Los resultados de este estudio tienen tanto implicaciones académicas como prácticas. Por un lado, confirman la validez del modelo de Fama-French para analizar portafolios europeos en periodos largos. Además, los gestores de inversiones que trabajan en entornos donde la transparencia y la comprensión de los riesgos son esenciales podrían preferir claramente el modelo lineal frente a alternativas más complejas, pero menos interpretables.

Por otro lado, los resultados ponen en duda la utilidad del uso de modelos de Machine Learning en las finanzas. Aunque estos modelos tienen mucho potencial, sobre todo cuando se trabaja con big data, su superioridad no está garantizada. Cuando el número de predictores es pequeño y las relaciones son lineales, el ML no solo no aporta una ventaja, sino que puede complicar el análisis sin mejorar los resultados.

En este sentido, el trabajo dice que los modelos de ML deberían aplicarse de solo en ciertas ocasiones donde haya muchos predictores potenciales, donde se sospecha de la presencia de fuertes no linealidades, donde existen interacciones complejas entre variables, o cuando la estructura temporal es inestable. Fuera de estos contextos, los modelos lineales siguen siendo herramientas mejores.

6.5. Limitaciones y líneas futuras de investigación

Como en todo análisis empírico, este trabajo presenta ciertas limitaciones que deben mencionarse. La primera viene del tamaño de la muestra y del número de variables empleadas, ya que, aunque 284 observaciones mensuales parecen una base de datos razonable, es posible que los modelos de ML necesiten una cantidad de datos más grande para poder desplegar todo su potencial.

Asimismo, el uso de un portafolio ponderado y relativamente sencillo puede limitar la detección de relaciones más complejas. De este modo, los estudios futuros podrían analizar carteras construidas a partir de estrategias más sofisticadas o incluso trabajar directamente con activos individuales, que tienen un grado de heterogeneidad es considerablemente mayor.

Del mismo modo, otra línea futura consiste en incorporar nuevos factores creados mediante ML, como por ejemplo, características textuales, señales alternativas,

indicadores macroeconómicos o datos ESG, con el objetivo de comprobar si efectivamente, el OLS sigue siendo mejor cuando aumenta la dimensionalidad del análisis.

Finalmente, podría ser interesante explorar modelos de ML diseñados específicamente para series temporales, como LSTM o Transformers, o métodos híbridos que combinen estructura económica con flexibilidad algorítmica.

7. CONCLUSIONES

El objetivo principal de este Trabajo de Fin de Grado ha sido analizar y comparar el funcionamiento del modelo de cinco factores de Fama y French con un modelo de Machine Learning, concretamente un Random Forest Regressor, aplicados a un portafolio formado por acciones europeas. De esta manera, se han evaluado ambos modelos utilizando una muestra amplia de datos mensuales que abarcan desde el 2002 al 2025, y se ha mantenido una metodología rigurosa y clara, además de revisar que los resultados fuesen coherentes desde una perspectiva económica.

Así, el estudio permite extraer tres conclusiones principales, que se detallan a continuación.

7.1. El modelo de Fama-French de cinco factores explica adecuadamente los rendimientos del portafolio europeo

En primer lugar, una de las conclusiones más relevantes es que el modelo lineal de cinco factores logra explicar una parte fundamental de la variación en los retornos mensuales del portafolio. Así, se observa que, el valor de R^2 obtenido en el conjunto de prueba, cercano al 0.67, demuestra que el modelo captura eficazmente las principales fuentes de variación sistemática. Asimismo, el hecho de que el término alpha sea prácticamente nulo indica que los factores empleados son más que suficientes para describir el comportamiento del portafolio, sin que queden anomalías o comportamientos sin una explicación consistente y coherente.

Además, estos resultados respaldan la fuerza del modelo propuesto por Fama y French y confirma que sigue siendo una herramienta sólida para analizar mercados desarrollados como el europeo, incluso considerando periodos de tiempo extensos y condiciones económicas diversas.

7.2. El modelo Random Forest no supera al modelo lineal, pese a su mayor complejidad

El segundo descubrimiento central del trabajo es que, al contrario de lo esperado, el modelo de Machine Learning no mejora la capacidad predictiva del modelo lineal. De hecho, las estadísticas del modelo, un R^2 de 0.61 y un MSE un poco superior, demuestran que el Random Forest es menos eficiente al predecir los retornos mensuales del portafolio.

Igualmente, este resultado es especialmente relevante, ya que muestra que la complejidad metodológica no siempre implica una mejora en el rendimiento. En general, los modelos de ML suelen ofrecer mejores resultados cuando los datos presentan relaciones no lineales, interacciones complejas o un número elevado de variables. Por el contrario, cuando los factores son escasos, están bien definidos teóricamente y representan relaciones lineales reales, el modelo clásico y un enfoque tradicional resulta más eficiente.

En este sentido, los resultados de este trabajo se alinean con gran parte de la literatura reciente que señala que los modelos lineales pueden ser sorprendentemente competitivos y difíciles de superar cuando el número de variables es reducido.

7.3. Implicaciones para la teoría, la práctica inversora y la investigación futura

Los resultados obtenidos ofrecen implicaciones relevantes tanto para la teoría del asset pricing como para la práctica financiera.

Centrándonos en lo teórico, el estudio confirma la solidez del marco de los cinco factores en el mercado europeo, lo que dice que este modelo captura gran parte de las dinámicas de riesgo relevantes. Además, se ve que los factores tradicionales siguen siendo útiles, a pesar de que en el contexto global sean las metodologías avanzadas de ML las que ganan protagonismo.

En cuanto a las implicaciones prácticas, los gestores de inversiones que deseen construir modelos transparentes, estables y basados en fundamentos económicos pueden encontrar en el modelo Fama-French una herramienta muy eficiente. Cabe destacar que, el hecho de que un modelo lineal supere a uno no lineal, en este contexto, es especialmente útil para quienes deben equilibrar precisión con interpretabilidad, como es el caso de las gestoras institucionales o los fondos regulados.

Finalmente, el estudio abre diversas líneas de investigación. En primer lugar, sería conveniente evaluar si la superioridad del modelo lineal sigue siendo así en carteras más heterogéneas o en datasets amplios que incluyan información contable detallada, indicadores macroeconómicos o datos derivados de técnicas textuales. En segundo lugar, podría considerarse la aplicación de modelos de ML diseñados específicamente para series temporales. De esta manera, se podría capturar dependencia temporal de forma más

eficiente. Por último, futuras investigaciones podrían contemplar la aplicación de métodos híbridos que combinen teoría económica con técnicas no lineales, con el objetivo de conseguir un equilibrio óptimo entre interpretabilidad y capacidad predictiva.

7.4. Síntesis final

En conjunto, este Trabajo de Fin de Grado demuestra que el modelo de los cinco factores de Fama y French sigue siendo muy útil para analizar los retornos de portafolios, incluso enfrentándose a alternativas más sofisticadas basadas en el Machine Learning. El análisis que se ha llevado a cabo confirma la relevancia de los factores tradicionales y su capacidad para capturar las principales fuentes de riesgo.

La comparación con el Random Forest muestra que la complejidad algorítmica no garantiza mejoras predictivas, especialmente cuando los factores están bien definidos y las relaciones son principalmente lineales. Este resultado da a entender que antes de recurrir a técnicas más avanzadas hay que comprender bien el contexto, los datos y los fundamentos teóricos del modelo.

Así, el trabajo concluye reafirmando la importancia de los modelos multifactoriales clásicos y ofreciendo una buena base para futuras investigaciones que exploren nuevos factores y nuevas técnicas que ayuden a entender los mercados financieros.

BIBLIOGRAFIA

Fama, E. F., & French, K. R. (1993). *Common risk factors in the returns on stocks and bonds. Journal of Financial Economics*, 33(1), 3–56.

Fama, E. F., & French, K. R. (2015). *A five-factor asset pricing model. Journal of Financial Economics*, 116(1), 1–22.

Fama, E. F., & French, K. R. (2019). *Comparing cross-section and time-series factor models*. (Working paper version, SSRN).

Hou, K., Xue, C., & Zhang, L. (2017). *Replicating anomalies* (NBER Working Paper No. 23394). National Bureau of Economic Research.

Chen, Y., Zhang, L., Xie, Z., Zhang, W., & Li, Q. (2023). *Unraveling asset pricing with AI: A systematic literature review.* (Working paper), Southwestern University of Finance and Economics.

ANEXO 1. CÓDIGO EN PYTHON

```
#
================================================================
=============

# INSTALACION Y LIBRERIAS

#
================================================================
==============

!pip install yfinance -q


import pandas as pd

import numpy as np

import yfinance as yf

import requests

import zipfile

import io

from datetime import datetime

import matplotlib.pyplot as plt

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean_squared_error, r2_score

import statsmodels.api as sm

import warnings

warnings.filterwarnings('ignore')
```

```
#
================================================================
==============
# PASO 1: CREAR DATAFRAME Y LIMPIEZA DE DATOS
#
================================================================
==============
# Descargamos los 5 factores que explican retornos de acciones
print("Descargando factores Fama-French...")


url =
"https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/Europe_5_Factors_CSV.zip"
response = requests.get(url)


# Extraer y leer el CSV
with zipfile.ZipFile(io.BytesIO(response.content)) as z:
    with z.open(z.namelist()[0]) as f:
        content = f.read().decode('utf-8', errors='ignore')


# Buscar donde empiezan los datos
lines = content.split('\n')
start = next(i for i, line in enumerate(lines) if 'Mkt-RF' in line)


# Extraer encabezado y datos
header = [h.strip() for h in lines[start].split(',')]
```

```
data = []
for line in lines[start+1:]:
    if not line.strip() or 'Annual' in line:
        break
    parts = [p.strip() for p in line.split(',')]
    if len(parts) == len(header) and parts[0].isdigit() and len(parts[0]) == 6:
        data.append(parts)

# Crear DataFrame
df_ff = pd.DataFrame(data, columns=header)
df_ff.columns = ['Date'] + list(df_ff.columns[1:])
df_ff['Date'] = pd.to_datetime(df_ff['Date'].str.strip(), format='%Y%m')
df_ff.set_index('Date', inplace=True)

# Convertir a numerico y a decimal
for col in df_ff.columns:
    df_ff[col] = pd.to_numeric(df_ff[col], errors='coerce')
df_ff = df_ff.dropna() / 100.0

# Renombrar columnas
mapping = {}
for col in df_ff.columns:
    c = col.upper().strip()
    if 'MKT' in c and 'RF' in c:
```

```
        mapping[col] = 'Mkt-RF'
    elif c in ['SMB', 'HML', 'RMW', 'CMA', 'RF']:
        mapping[col] = c

df_ff = df_ff.rename(columns=mapping)[['Mkt-RF', 'SMB', 'HML', 'RMW', 'CMA',
'RF']]
df_ff.index = df_ff.index.tz_localize(None) + pd.offsets.MonthEnd(0)

print(f"OK. Periodo: {df_ff.index.min():%Y-%m} a {df_ff.index.max():%Y-%m}
({len(df_ff)} meses)")

#
==========================================================================
==============
# PASO 2: DESCARGAR PRECIOS DE ACCIONES
#
==========================================================================
==============
# Descargamos precios de acciones europeas para crear un portafolio
print("\nDescargando precios de acciones...")

tickers = ['SAN.MC', 'BBVA.MC', 'ITX.MC', 'IBE.MC', 'TEF.MC', 'REP.MC',
'ACS.MC',
           'MC.PA', 'OR.PA', 'AI.PA', 'BNP.PA', 'SIE.DE', 'SAP.DE', 'SHEL.L', 'AZN.L']

start_date = df_ff.index.min() - pd.DateOffset(months=3)
```

```
prices_list = []

for ticker in tickers:
    try:
        data = yf.Ticker(ticker).history(start=start_date, end=datetime.now(),
auto_adjust=True)
        if len(data) > 100 and data['Close'].notna().sum() / len(data) > 0.7:
            prices_list.append(data['Close'].rename(ticker))
            print(f"  {ticker}: OK")
            if len(prices_list) >= 7:
                break
    except:
        print(f"  {ticker}: ERROR")

# Combinar precios
prices = pd.concat(prices_list, axis=1).dropna(how='all')
prices.index = prices.index.tz_localize(None)
prices = prices.dropna(thresh=int(len(prices)*0.6), axis=1)

print(f"\nAcciones descargadas: {list(prices.columns)}")

#
==========================================================================
==============
# PASO 3: CALCULAR RETORNOS MENSUALES
```

```
#
================================================================
==============

# Convertimos precios diarios en retornos mensuales

print("\nCalculando retornos mensuales...")


prices_monthly = prices.resample('M').last()

returns_monthly = np.log(prices_monthly / prices_monthly.shift(1)).dropna()

portfolio_return = returns_monthly.mean(axis=1)  # Promedio de todas las acciones

portfolio_return.index = portfolio_return.index.tz_localize(None) +
pd.offsets.MonthEnd(0)


print(f"Retornos calculados: {len(portfolio_return)} meses")


#
================================================================
==============

# PASO 4: COMBINAR DATOS

#
================================================================
==============

# Juntamos factores FF con retornos del portafolio

print("\nCombinando datos...")


df_final = df_ff.merge(

    pd.DataFrame({'Portfolio_Return': portfolio_return}),
```

```
    left_index=True,
    right_index=True,
    how='inner'
).dropna()

print(f"Dataset final: {len(df_final)} observaciones")
print(f"Periodo: {df_final.index.min():%Y-%m} a {df_final.index.max():%Y-%m}\n")

#
==========================================================================
==============
# PASO 5: DIVIDIR DATOS
#
==========================================================================
==============
# Separamos variables explicativas (X) de objetivo (y)
# y dividimos en entrenamiento (80%) y prueba (20%)

X = df_final[['Mkt-RF', 'SMB', 'HML', 'RMW', 'CMA']]
y = df_final['Portfolio_Return']

split = int(len(df_final) * 0.8)
X_train, X_test = X.iloc[:split], X.iloc[split:]
y_train, y_test = y.iloc[:split], y.iloc[split:]

print(f"Entrenamiento: {len(X_train)} meses")
```

```
print(f"Prueba: {len(X_test)} meses\n")


#
==========================================================================
==============
# PASO 6: MODELO OLS (FAMA-FRENCH)
#
==========================================================================
==============
# Modelo clasico de regresion lineal
print("Entrenando modelo OLS...")


model_ols = sm.OLS(y_train, sm.add_constant(X_train)).fit()
y_pred_ols = model_ols.predict(sm.add_constant(X_test))


r2_ols = r2_score(y_test, y_pred_ols)
mse_ols = mean_squared_error(y_test, y_pred_ols)


print(f"R2: {r2_ols:.4f}")
print(f"MSE: {mse_ols:.6f}\n")


#
==========================================================================
==============
# PASO 7: MODELO RANDOM FOREST
```

```
#
==========================================================================
==============
# Modelo de Machine Learning
print("Entrenando Random Forest...")

rf_model = RandomForestRegressor(
    n_estimators=300,
    max_depth=8,
    min_samples_leaf=5,
    max_features='sqrt',
    random_state=42,
    n_jobs=-1
)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)

r2_rf = r2_score(y_test, y_pred_rf)
mse_rf = mean_squared_error(y_test, y_pred_rf)

print(f"R2: {r2_rf:.4f}")
print(f"MSE: {mse_rf:.6f}\n")

# Importancia de factores
importance = pd.DataFrame({
```

```
    'Factor': X.columns,
    'Importancia': rf_model.feature_importances_
}).sort_values('Importancia', ascending=False)

print("Importancia de factores:")
print(importance.to_string(index=False))

#
==========
# PASO 8: COMPARACION
#
==========
print("\n" + "="*60)
print("COMPARACION DE MODELOS")
print("="*60)

comparison = pd.DataFrame({
    'Modelo': ['OLS', 'Random Forest'],
    'R2': [r2_ols, r2_rf],
    'MSE': [mse_ols, mse_rf]
})
print(comparison.to_string(index=False))

if r2_rf > r2_ols and mse_rf < mse_ols:
```

```
    print("\nRandom Forest es MEJOR")
elif r2_ols > r2_rf and mse_ols < mse_rf:
    print("\nOLS es MEJOR")
else:
    print("\nRendimiento SIMILAR")

#
==========================================================================
==============
# PASO 9: VISUALIZACIONES
#
==========================================================================
==============
print("\nGenerando graficos...")

fig = plt.figure(figsize=(16, 10))

# Grafico 1: Importancia
ax1 = plt.subplot(2, 2, 1)
ax1.barh(importance['Factor'], importance['Importancia'], color='steelblue')
ax1.set_xlabel('Importancia')
ax1.set_title('Importancia de Factores (RF)', fontweight='bold')
ax1.invert_yaxis()

# Grafico 2: Coeficientes OLS
ax2 = plt.subplot(2, 2, 2)
```

```
coef = model_ols.params[1:]
colors = ['green' if x > 0 else 'red' for x in coef]
ax2.barh(coef.index, coef.values, color=colors, alpha=0.7)
ax2.axvline(x=0, color='black', linestyle='--')
ax2.set_xlabel('Coeficiente Beta')
ax2.set_title('Coeficientes Beta (OLS)', fontweight='bold')
ax2.invert_yaxis()

# Grafico 3: Serie temporal
ax3 = plt.subplot(2, 1, 2)
ax3.plot(y_test.index, y_test.values, 'ko-', label='Real', linewidth=2)
ax3.plot(y_test.index, y_pred_ols, 'b--s', label='OLS', linewidth=2)
ax3.plot(y_test.index, y_pred_rf, 'r:^', label='RF', linewidth=2)
ax3.set_xlabel('Fecha')
ax3.set_ylabel('Retorno')
ax3.set_title('Predicciones vs Reales', fontweight='bold')
ax3.legend()
ax3.grid(True, alpha=0.3)
ax3.axhline(y=0, color='gray', linestyle='-', linewidth=0.5)
plt.setp(ax3.xaxis.get_majorticklabels(), rotation=45)

plt.tight_layout()
plt.savefig('resultados.png', dpi=300, bbox_inches='tight')
print("Grafico guardado: resultados.png")
```

```
plt.show()


# Scatter plots
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(14, 5))

ax1.scatter(y_test, y_pred_ols, alpha=0.6, edgecolors='black')
ax1.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--', lw=2)
ax1.set_xlabel('Real')
ax1.set_ylabel('OLS')
ax1.set_title(f'OLS (R2={r2_ols:.4f})', fontweight='bold')
ax1.grid(True, alpha=0.3)

ax2.scatter(y_test, y_pred_rf, alpha=0.6, color='red', edgecolors='black')
ax2.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'k--', lw=2)
ax2.set_xlabel('Real')
ax2.set_ylabel('RF')
ax2.set_title(f'Random Forest (R2={r2_rf:.4f})', fontweight='bold')
ax2.grid(True, alpha=0.3)

plt.tight_layout()
plt.savefig('scatter.png', dpi=300, bbox_inches='tight')
print("Grafico guardado: scatter.png")
plt.show()
```

```
#
==========================================================================
==============
# RESUMEN FINAL
#
==========================================================================
==============
print("\n" + "="*70)
print("RESUMEN EJECUTIVO DEL ANALISIS")
print("="*70)

print(f"""
DATOS PROCESADOS:
  - Factores: Fama-French 5 (Mkt-RF, SMB, HML, RMW, CMA)
  - Mercado: Europa
  - Periodo: {df_final.index.min():%Y-%m} a {df_final.index.max():%Y-%m}
  - Observaciones: {len(df_final)} meses
  - Acciones: {list(prices.columns)}
  - Division: {len(X_train)} meses entrenamiento / {len(X_test)} meses prueba

RESULTADOS MODELO OLS (FAMA-FRENCH):
  - R-cuadrado: {r2_ols:.4f} ({r2_ols*100:.2f}% varianza explicada)
  - MSE: {mse_ols:.6f}
  - Alpha: {model_ols.params['const']:.6f}
  - Interpretacion: Modelo clasico que asume relaciones LINEALES
    entre factores y retornos
```

RESULTADOS RANDOM FOREST:

- R-cuadrado: {r2_rf:.4f} ({r2_rf*100:.2f}% varianza explicada)

- MSE: {mse_rf:.6f}

- Factor clave: {importance.iloc[0]['Factor']} ({importance.iloc[0]['Importancia']:.3f})

- Interpretacion: Modelo ML que captura relaciones NO LINEALES

e interacciones entre factores

COMPARACION Y CONCLUSION:

- Diferencia R2: {abs(r2_rf - r2_ols):.4f} ({"RF mejor" if r2_rf > r2_ols else "OLS mejor" if r2_ols > r2_rf else "Similar"})

- Diferencia MSE: {abs(mse_rf - mse_ols):.6f} ({"RF mejor" if mse_rf < mse_ols else "OLS mejor" if mse_ols < mse_rf else "Similar"})

INTERPRETACION FINAL:

""")

Generar interpretacion detallada segun resultados

if r2_rf > r2_ols and mse_rf < mse_ols:

print(f""" El Random Forest supera al OLS con un R2 {abs(r2_rf - r2_ols):.4f} puntos mayor

y un error {abs(mse_ols - mse_rf):.6f} puntos menor.

Esto sugiere que:

1. Existen relaciones NO LINEALES entre los factores y los retornos

```
  2. Hay INTERACCIONES entre factores que el modelo lineal no captura
  3. Los modelos de ML pueden aportar valor en prediccion financiera

  Implicacion: Para este mercado y periodo, la complejidad del RF
  esta justificada y mejora la capacidad predictiva.""")

elif r2_ols > r2_rf and mse_ols < mse_rf:
    print(f"""  El OLS supera al Random Forest con un R2 {abs(r2_ols - r2_rf):.4f}
puntos mayor
  y un error {abs(mse_rf - mse_ols):.6f} puntos menor.

  Esto sugiere que:
  1. Las relaciones son predominantemente LINEALES
  2. El modelo academico clasico es SUFICIENTE para este mercado
  3. La complejidad adicional del RF no aporta valor (posible overfitting)

  Implicacion: El modelo de Fama-French se valida empiricamente
  y la simplicidad del OLS es preferible (interpretabilidad + precision).""")

elif abs(r2_rf - r2_ols) < 0.02 and abs(mse_rf - mse_ols) < 0.0001:
    print(f"""  Ambos modelos tienen rendimiento MUY SIMILAR:
  - Diferencia R2: {abs(r2_rf - r2_ols):.4f} (practicamente identica)
  - Diferencia MSE: {abs(mse_rf - mse_ols):.6f} (insignificante)

  Esto sugiere que:
```

```
  1. Las relaciones son principalmente lineales
  2. No hay suficientes datos para que RF aproveche su flexibilidad
  3. La complejidad del RF no esta justificada

  Implicacion: En la practica, se preferiria OLS por su SIMPLICIDAD,
  INTERPRETABILIDAD y menor coste computacional, dado que la precision
  es equivalente.""")

else:
    mejor_r2 = "RF" if r2_rf > r2_ols else "OLS"
    mejor_mse = "RF" if mse_rf < mse_ols else "OLS"
    print(f"""  Los modelos tienen trade-offs:
  - Mejor R2: {mejor_r2} ({r2_rf:.4f} vs {r2_ols:.4f})
  - Mejor MSE: {mejor_mse} ({mse_rf:.6f} vs {mse_ols:.6f})

  Esto sugiere que:
  1. Ningun modelo domina claramente al otro
  2. Puede haber overfitting en alguno de los modelos
  3. El resultado depende de la metrica priorizada

  Implicacion: Se recomienda validacion adicional con mas datos
  o usar ensemble de ambos modelos para combinar sus fortalezas.""")

print("\n" + "="*70)
```

```
print("ANALISIS COMPLETADO")

print("="*70)
```

ANEXO 2. GRÁFICOS DE PYTHON

Gráfico 1:

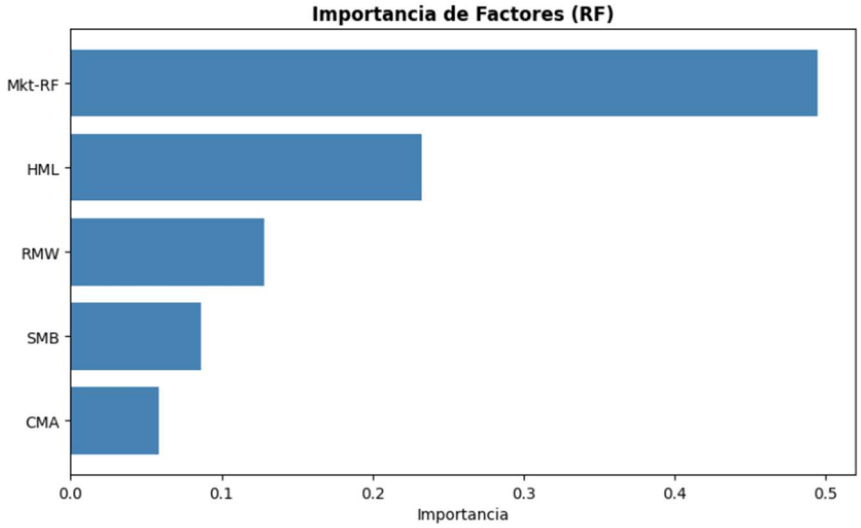

Gráfico 2:

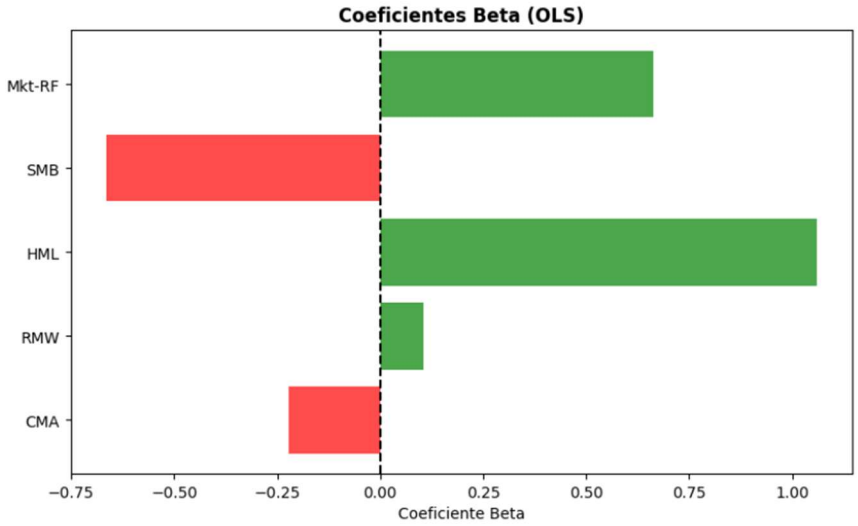

Gráfico 3:

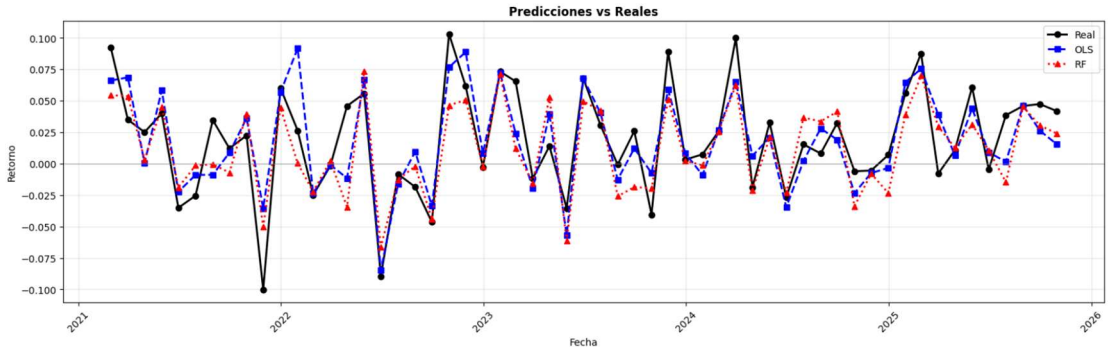

Gráfico 4:

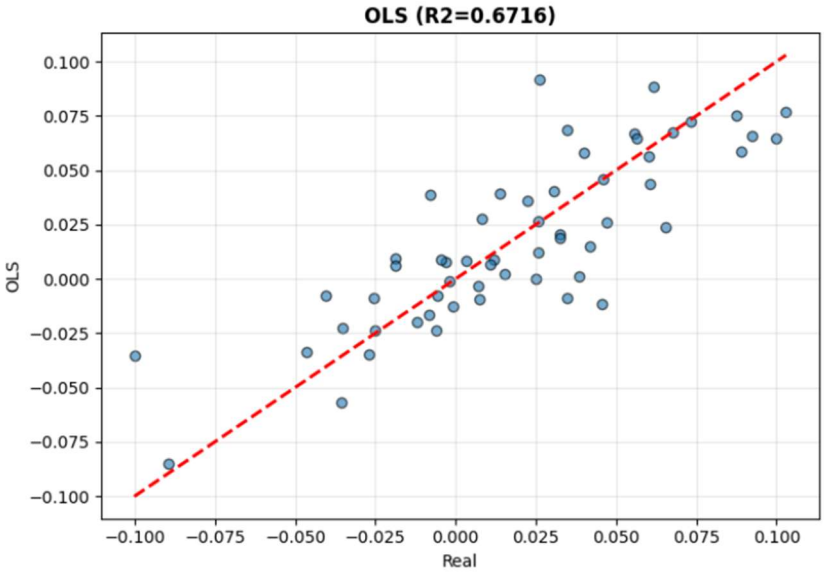

Gráfico 5: