Investigación en Mercados Financieros en la Era de la Inteligencia Artificial y el Big Data*

Roberto Pascual[1] y José Penalva[2]

[1]Universidad de las Islas Baleares
[2]Universidad Carlos III de Madrid

Julio 2024

1. Introducción

Vivimos tiempos de cambios tecnológicos acelerados, y éstos están afectando todos los aspectos de la sociedad. La investigación académica no es una excepción. En este trabajo queremos dar cuenta de cómo el progreso tecnológico está transformando la investigación en el ámbito de los mercados financieros, así como identificar los desafíos y oportunidades de futuro que este proceso genera.

Este capítulo no pretende ser un estudio del estado de la literatura, ni un análisis detallado de la extensa librería de nuevos métodos y técnicas disponibles para el investigador. Se trata de una visión personal, subjetiva en su elección de temas y referencias, que esperamos sea interesante y útil para el lector.

Ciertos conceptos, como "*Inteligencia Artificial*", *"Big Data"* y "*Machine Learning*", por ser clave en nuestra discusión, aparecerán repetidamente en las siguientes secciones. Consideramos importante empezar aclarando qué entendemos por cada uno de ellos cuando los aplicamos a la investigación en Mercados Financieros.

En la Sección 2 analizamos cómo las nuevas tecnologías han afectado el insumo en el proyecto de investigación: los datos. Utilizamos el término "datos" como un genérico para referirnos a un input objetivo y sistematizado sobre la cual se construye la investigación financiera. Un elemento clave para nosotros es que

*Los autores desean agradecer a Álvaro Arroyo, Faycal Drissi, Fernando Moreno-Pino, Daniel Peña, y Harrison Waldon por sus comentarios y sugerencias.

este input puede (al menos potencialmente) ser utilizado por otros investigadores tanto para replicar como para ampliar la investigación inicial. La culminación de la llamada era digital o era de la información ha traído consigo un aumento del volumen, calidad, y diversidad de fuentes de datos potencialmente valiosos para los investigadores, un fenómeno que se conoce como *"Big Data"* (BD).

¿Qué significa exactamente BD en Finanzas? Goldstein *et al.* (2021) sostienen que la tradicional acepción basada en las tres "V" (volumen, velocidad, y variedad), que proviene de ciencias como la Ingeniería, la Estadística, o la Informática, no refleja ni las oportunidades ni los retos que el BD representa para la investigación y la práctica de las Finanzas.[1] En su lugar, proponen una definición aplicada a la Economía Financiera a la que nosotros nos adherimos. Según esta definición, el BD en Finanzas tiene que tener tres propiedades: gran tamaño, alta dimensión, y estructura compleja.

La primera de estas propiedades, "gran tamaño", hace referencia al volumen de datos. El ejemplo más palpable de datos de gran volumen en Finanzas son los datos de alta frecuencia que se utilizan en la investigación de Microestructura de los Mercados Financieros (por ej., Hasbrouck, 2007; Foucault *et al.*, 2024). En la Sección 2.1 hablaremos de las características y complejidades asociadas a trabajar con estos datos de alta frecuencia. Por "alta dimensión" nos referimos a situaciones en que los datos contienen muchas variables respecto al tamaño de la muestra. Un ejemplo de alta dimensión sería el "zoo de factores" en la investigación sobre valoración de activos (p. ej., Harvey *et al.*, 2016; Giglio *et al.*, 2021).[2] Bajo el mismo paraguas también incorporamos la riqueza de las interacciones (no-lineales) entre las variables que aumentan la complejidad en las relaciones entre ellas. Finalmente, por "estructura compleja" entendemos datos no estructurados, es decir, que no vienen compartimentados por filas y columnas, sino que siguen formatos menos tradicionales como texto, imágenes, vídeo, o audio.

Los datos no estructurados proporcionan valor añadido al investigador en tanto en cuanto permiten extraer información útil que no puede derivarse de los datos estructurados. En años recientes, proliferan estudios en Economía Financiera que utilizan estos datos para extraer indicadores de sentimiento (p. ej., Tetlock, 2007; Obaid *et al.*, 2022), medidas del nivel de atención del inversor (p. ej., Da *et al.*, 2015), indicadores de cultura empresarial (Li *et al.*, 2021), temáticas de un texto (p. ej., Fedyk, 2023), y patrones complejos en gráficos de series de precios (p. ej., Jiang *et al.*, 2023), entre otras aplicaciones. Hablaremos de los datos no estructurados aplicados a la investigación en Mercados Financieros en la Sección 2.2.

[1] "Volumen" hace referencia a la cantidad de datos, "velocidad" a la rapidez con que los datos de crean y almacenan, y "variedad" a la heterogeneidad de los datos, tanto en cuanto a su tipología como a su formato. Algunos añaden una cuarta "V" de "veracidad", que hace referencia a la calidad y precisión de los datos.

[2] El término "*factor zoo*" fue introducido por Cochrane (2011) para referirse al creciente número de factores utilizados en modelos de valoración, y que se cuentan por cientos.

La mayor presencia del BD en la investigación en Finanzas especialmente en la última década ha sido posible gracias no sólo a la cada vez mayor disponibilidad de datos masivos, sino también al desarrollo tecnológico que permite almacenar, gestionar, y analizar grandes volúmenes de datos de forma cada vez más eficiente. Además, el desarrollo de la inteligencia artificial (IA) ha traído consigo nuevas metodologías para lidiar con el problema de la alta dimensión, o para procesar datos no estructurados.

Por IA entendemos la simulación de procesos asociados con la inteligencia humana por parte de sistemas informáticos. En otras palabras, el objetivo de la IA es crear sistemas capaces de emular el razonamiento humano.[3] Esos sistemas normalmente implican una fase de aprendizaje (la máquina necesita conocer el input que debe procesar y disponer de unas reglas que le indiquen cómo usar ese input), razonamiento (utilizar esos datos y reglas para llegar a conclusiones y/o recomendaciones) y retroalimentación (aprender de los errores y aciertos propios).[4]

El pionero de la IA, Alan Turing (1912-1954), en su artículo "*Computing Machinery and Intelligence*" publicado en la revista *Mind* en 1950 ya anticipa la dificultad de desarrollar una máquina capaz de pensar como un humano adulto por la mera programación, y describe la necesidad de que la máquina aprenda por sí misma, de la misma forma que un niño se educa para alcanzar un cerebro adulto. La idea de que en lugar de que expertos impongan reglas a la máquina sea la máquina la que de forma automática aprenda esas reglas es la base en la que se fundamenta el *machine learning* (ML) (Muggleton, 2014).

El aprendizaje automático o ML es una rama de la IA centrada en desarrollar algoritmos y modelos estadísticos que permiten a la máquina aprender por sí misma. A partir de unos datos de entrenamiento, estos algoritmos detectan patrones y tendencias que luego aplican a nuevos datos para realizar predicciones o tomar decisiones. El ML involucra diferentes técnicas de aprendizaje (supervisado y no-supervisado) que difieren en la naturaleza de la retroalimentación proporcionada durante la fase de entrenamiento del algoritmo.

En el aprendizaje supervisado (*supervised learning*) el algoritmo aprende a partir de datos de entrenamiento etiquetados, en forma de pares de entrada y salida. El objetivo del algoritmo es derivar una correspondencia de las entradas a las salidas que pueda generalizarse a datos fuera de la muestra minimizando el error de predicción. La retroalimentación es explícita, ya que el algoritmo conoce la verdadera salida asociada a cada entrada. Por el contrario, en el aprendizaje no supervisado (*un-supervised learning*) el algoritmo recibe datos de entrada pero no de salida y tiene que descubrir patrones, estructuras, o representaciones

[3] El término "IA" se atribuye a John McCarthy (1927-2011) que la define como "la ciencia y la ingeniería de hacer máquinas inteligentes".

[4] Utilizamos el lenguaje común en esta literatura, aunque con ello no queremos decir que las máquinas puedan de una manera real aprender o razonar tal y como entendemos que hace la mente humana.

subyacentes dentro de dichos datos. La retroalimentación es implícita, y suele implicar la evaluación de la calidad de los patrones descubiertos. En el aprendizaje por refuerzo (*reinforcement learning*), el algoritmo interacciona con un entorno en el que toma decisiones y que le proporciona retroalimentación en la forma de recompensas y penalizaciones. En este caso, el objetivo es buscar la estrategia que maximice la recompensa acumulada a lo largo del tiempo.

Además de la creciente disponibilidad de BD, Kelly y Xiu (2023) destacan otras dos características de la investigación en Finanzas que la hacen suelo fértil para las técnicas de IA: la existencia de un conjunto de información condicional muy amplio y unas formas funcionales ambiguas. Nosotros añadiríamos una tercera característica: un alto cociente de ruido sobre información en los precios, que aumenta con la frecuencia de los mismos.

Consideremos el caso de la valoración de activos. Idealmente, el investigador quisiera incorporar como input de su modelo de valoración toda aquella información que resulte relevante a los agentes del mercado a la hora de formar sus expectativas sobre el verdadero valor de un activo financiero. El problema es que la variedad de fuentes de información en las que basan sus expectativas es cada vez más amplia (Martin y Nagel, 2022). Los inversores pueden acceder en red a informes de empresas, informes económicos, u opiniones de expertos (expresadas en redes sociales) para realizar valoraciones. También producen informes los proveedores de datos (como Bloomberg o Reuters) y "Fintechs" (como iSentium, Dataminr, o Eagle Alpha) que utilizan "*news analytics*" para extraer señales a partir de datos no estructurados (noticias, comunicados de prensa, anuncios del mercado de valores, tweets, imágenes satelitales, etc.) que luego venden a los inversores para que los usen como inputs en sus algoritmos de negociación (Dugast y Foucault, 2018). Además, hay factores de valoración que afectan a la evolución a lo largo tiempo del precio de un activo, y factores que explican diferencias en el nivel de precio entre activos. Cochrane (2009) señala que toda esa información condicional ni siquiera es observable para el investigador, y aunque lo fuese, los modelos econométricos tradicionales serían incapaces de incorporarlos.

Por otro lado, a la hora de definir el modelo de predicción, no existe un consenso sobre cuáles son las formas funcionales más apropiadas, tanto en modelos estructurales como en modelos en forma reducida. Al fin y al cabo, los inversores utilizan e interpretan la información de formas complejas, no directamente observables por el investigador. Por tanto, la crítica de Cochrane (2009) puede extenderse también a la elección de la forma funcional.

Por último, las series de precios de activos financieros incorporan una gran cantidad de ruido por la naturaleza misma de su proceso de formación. Los precios de los activos contienen un componente de largo plazo, que cambia a medida que se revisa el consenso sobre el verdadero valor del activo como respuesta a la llegada de nueva información (pública o privada) al mercado, y un componente de ruido causado por divergencias de opinión, errores humanos no-

sistemáticos, fricciones reales (competencia imperfecta, gestión de inventario, costes operativos etc.), y otras fluctuaciones.

El ML ofrece una serie de procedimientos que permiten al investigador lidiar con problemas que involucran muchas variables, al tiempo que le ofrece la posibilidad de llevar a cabo un análisis estadístico aún cuando desconoce de forma cierta la forma funcional del modelo, y ajustar por exceso de ruido y relaciones espúreas en los datos. De hecho, la forma habitual en que estos algoritmos operan es considerar una amplia colección de posibles especificaciones, que incluyen no linealidades e interacciones entre las variables, y apoyarse en los datos para elegir el modelo más efectivo, proceso que se conoce como *model tuning*. En este sentido, el ML está más orientado hacia maximizar la precisión en la predicción que hacia la inferencia estadística. El ML prioriza modelos simples con un buen desempeño fuera de muestra sobre modelos complejos que ajusten muy bien dentro de la muestra pero con un desempeño fuera de ella muy pobre. Los procesos de ML incluyen mecanismos de regularización que restringen el tamaño del modelo para evitar problemas de sobreajuste (*overfitting*), en los que un modelo sobre-dimensionado o innecesariamente complejo se ajusta al ruido inherente en los datos en lugar de captar la relación económica de interés subyacente. El modelo óptimo en ML sería aquel lo suficientemente grande como para detectar de manera fiable relaciones predictivas potencialmente complejas en los datos, pero no tan flexible como para estar dominado por el *overfitting* y sufrir fuera de muestra (Gu *et al.*, 2020a).

Además de ampliar el abanico de técnicas y herramientas para la investigación empírica, ¿puede el ML mejorar la teoría financiera? Una crítica habitual al ML es, precisamente, la carencia de una asociación explícita con la teoría económica. [5] Cabe entender que el ML parte del objetivo de mejorar nuestra capacidad predictiva por encima de cualquier otra cosa. No obstante, es precisamente a raíz de las nuevas realidades empíricas reveladas que el ML puede incentivar el desarrollo de nuevas teorías que nos permitan entender mejor los mecanismos económicos subyacentes a los modelos y los resultados que se derivan de la aplicación de estas técnicas.

En la Sección 3 ofrecemos una revisión de las metodologías de ML más utilizadas hasta la fecha en la investigación sobre Mercados Financieros, y seleccionamos aplicaciones ilustrativas de algunas de ellas. Entre otras, veremos ejemplos de métodos basados en redes neuronales, como *convolutional networks* que se aplican al análisis de imágenes; métodos de regularización como *ElasticNet*; técnicas para lidiar con el problema de la alta dimensión, como *principal components regression* y *partial least squares*; métodos basados en árboles de decisión como *random forest*; y algoritmos para el análisis lingüístico, como *encoders*, entre otros. Finalmente, en la Sección 4, concluimos.

[5] Como veremos esta carencia no es universal, ya que hay trabajos que conectan los métodos de ML con la teoría matemática (Ledoit y Wolf, 2020), y con la teoría financiera (Kelly *et al.*, 2024).

2. Big Data y Mercados Financieros

> *Nuestra capacidad para analizar el funcionamiento de los mercados financieros está limitada por la disponibilidad de datos relevantes*
>
> Goodhart y O'Hara (1997)

> *La Economía Financiera es, sin duda, la más empírica de todas las ciencias sociales. Esto se debe, en parte, a su orientación deliberadamente práctica y a la disponibilidad de datos de alta calidad sobre los mercados financieros.*
>
> Andersen (2000)

La revolución de la IA y del BD ha cambiado la industria financiera. Por un lado, existe una carrera tecnológica entre las empresas de inversión para conseguir tecnología que sea sólo un poco más rápida o inteligente que la de sus competidores.[6] Por otro lado, existe evidencia de que los gestores de carteras están cambiando estrategias discrecionales basadas en el juicio humano y el análisis fundamental por estrategias cuantitativas basadas en el análisis algorítmico de datos (p. ej., Abis, 2017). Así, donde antes la negociación en los mercados financieros estaba en manos de los MBAs, ahora está dominada por los "*quants*" (Zuckerman y Hope, 2017).

Estos analistas cuantitativos utilizan potentes ordenadores, así como complejos modelos matemáticos y estadísticos, para procesar simultáneamente billones de Gbytes de datos procedentes de diferentes fuentes y variedad de formatos (i.e., BD). Con estos modelos valoran activos, gestionan riesgos, y detectan oportunidades de negociación beneficiosas, permitiendo incluso que sus algoritmos aprendan de forma continuada de sus propios aciertos y errores, en lo que se conoce como *Artificially Intelligent Algorithmic Trading* (AI-AT).[7] Incluso los órganos de supervisión están contratando *quants* para poder entender una tecnología que cambia en tiempo real y a gran velocidad, así como gestionar el BD que ellos mismos recogen.[8]

Posiblemente, el caso más conocido de *quants* en la industria financiera y que ha generado (por polémico) mayor interés entre los reguladores, los medios de comunicación, y los académicos sea el de la negociación de alta frecuencia (HFT)

[6] Véase, por ejemplo, el caso de JP Morgan Bloomberg News (2019).

[7] Véase Marr (2019).

[8] Véase el discurso del Economista Jefe de la *Securities Exchange Comission* (SEC), S.P. Kothary, en el NBER-RFS *Summer Conference in Big Data* U.S. Securities and Exchange Commission (2020).

(p. ej., Cartea *et al.*, 2015). Los algoritmos de HFT explotan oportunidades de beneficio cuyas duraciones se miden en unidades de tiempo inferiores al segundo.[9] Para ello, sus algoritmos deben acceder, procesar y reaccionar a señales de diferente naturaleza a velocidades extremas. Estas señales pueden proceder de la negociación (transacciones, cotizaciones) (p. ej., Kwan *et al.*, 2024), anuncios públicos (p. ej., Chordia *et al.*, 2018), pero también de fuentes menos tradicionales, como mensajes de X/Twitter (p. ej., Dugast y Foucault, 2018). Pero el mundo del análisis cuantitativo va mas allá del caso del HFT, abarcando estrategias de optimización de carteras, minimización de costes, gestión de riesgos, predicción de rentabilidades, etc. con horizontes de inversión que pueden ir de días a años. El peso de toda esta negociación algorítmica (AT) en el volumen de negociación (*turnover*) se estima en un 65 % en EE.UU., 45 % en Europa, y 38 % en Asia en 2017.[10]

Esta revolución está también empezando a afectar de forma significativa a la investigación académica. El desarrollo tecnológico ha incrementado enormemente la tipología, volumen, y calidad de los datos disponibles para los investigadores (tanto académicos como no académicos), como demuestra el crecimiento en el sector de la información económica.[11] Como consecuencia, es cada vez más común encontrar en las revistas de referencia en Mercados Financieros trabajos que utilizan BD y técnicas de ML.

En la Figura 1 mostramos el número de publicaciones en revistas científicas generalistas sobre Finanzas vinculadas a una o varias de las siguientes temáticas: IA, BD, o ML. Incluimos las tres revistas de mayor índice de impacto (*Journal of Finance*, *Review of Financial Studies* y *Journal of Financial Economics*), así como otras tres revistas de referencia en cuanto a estudios sobre mercados financieros se trata (*Journal of Financial and Quantitative Analysis*, *Review of Finance*, y *Journal of Financial Markets*). Puede observarse como gran parte de la producción científica sobre éstas tres temáticas se concentra mayormente en las tres revistas más importantes, lo que sugiere que estamos ante un tópico que despierta un gran interés, pero que está todavía en un estadio incipiente. De hecho, en la Figura 2 podemos observar cómo una gran mayoría de estos trabajos se ha publicado en los últimos seis años.

[Figura 1]

[Figura 2]

[9] Véase Budish *et al.* (2015), Baron *et al.* (2019), Shkilko y Sokolov (2020), Aquilina *et al.* (2022)

[10] Fuente: Hong Kong Institute for Monetary and Financial Research (HKIMR) (2021).

[11] El número de empleados en el sector NAICS 519 (portales de búsqueda en la web, bibliotecas, archivos y otros servicios de información) ha crecido casi 5 veces en los EE.UU. durante los últimos 30 años, pasando de 37,000 en 1994 a 177,600 a finales de 2013 (Fuente: Bureau of Labor Statistics)

En las Figura 3 mostramos similares estadísticas pero para cuatro revistas especializadas en Finanzas Cuantitativas y Matemáticas Financieras. Podemos observar cómo la revista *Quantitative Finance* se ha convertido en un nicho para la publicación de artículos vinculados a IA, BD, o ML. Sin embargo, éstas revistas tienen índices de impacto significativamente inferiores que las de la Figura 3.

[Figura 3]

2.1. Datos estructurados

Proveedores de datos como Bloomberg, Reuters, Eikon, BMLL, FactSet, InfoTrie, FinnWorlds, TradeFeeds, Barchart, WRDS, etc. proporcionan información estructurada a diferentes frecuencias para decenas de mercados y miles de instrumentos financieros. Bases de datos como Datastream, Compustat, CRSP, ORBIS, I/B/E/S y GFD ofrecen información histórica y periódica sobre empresas cotizadas incluyendo cuentas de resultados, balances, flujos de caja, anuncios de beneficios, previsiones de analistas etc. así como información diaria sobre activos financieros (precio de apertura, cierre, máximo, mínimo, capitalización bursátil, volumen negociado etc.), abarcando más del 90 % de la capitalización bursátil mundial.

Mención aparte merecen las bases de datos de alta frecuencia o *tick data*. Por "datos de alta frecuencia" entendemos datos granulares sobre eventos discretos recopilados y ordenados en tiempo continuo, sin agregación temporal.[12] Ejemplos de estas bases de datos incluyen NYSE Trade and Quote (TAQ), Nasdaq Equity Tick History, Nasdaq TotalView-ITCH, Eikon, LSEG Tick History (anteriormente llamado TRTH), EUROFIDAI High-Frequency Database, y BME Tick Data.

Los acontecimientos recogidos en estas bases de datos pueden ser de diferente naturaleza:

(a) Transacciones (datos de "Nivel1"): Una transacción resulta del cruce de una oferta de compra y otra de venta compatibles que llegan secuencialmente al mercado. Para cada transacción es común encontrar información sobre el precio, volumen ("tamaño") y tiempo de ejecución. Es menos habitual conocer el signo (compra/venta) de la operación, que viene determinado por quién es el tomador de liquidez (i.e., la parte agresiva) en cada operación. Éste puede inferirse, no obstante, utilizando algoritmos de clasificación (Chakrabarty *et al.*, 2015), o emparejando el registro de transacciones con la evolución del libro de órdenes (Pascual y Veredas, 2010).

[12] Hussain *et al.* (2023) identifican 2920 artículos científicos publicados que utilizan datos de alta frecuencia entre 1977 y 2019.

Como ejemplos, Easley *et al.* (1996) proponen una metodología para estimar la probabilidad de negociación informada de un activo (PIN) que requiere de datos de Nivel1. Huang y Stoll (1997) utilizan datos de Nivel1 para estimar los componentes teóricos de la horquilla de precios o *bid-ask spread*. Más recientemente, Brogaard *et al.* (2014) usan datos de Nivel1 del Nasdaq para estudiar la contribución de las transacciones iniciadas por los *traders* de alta frecuencia (HFTs) en la formación del precio.

(b) Cotizaciones (datos de "Nivel2"): Las cotizaciones reflejan las ofertas de compra y venta reveladas y disponibles para negociar en cada instante. Para cada cotización, encontramos un precio al que se está dispuesto a negociar, un volumen máximo ofrecido a dicho precio ("profundidad"), y el momento del tiempo en que se registra. La gran mayoría de bases de datos de alta frecuencia para mercados bursátiles ofrecen datos sobre las mejores ofertas de compra (*bid*) y venta (*ask*) (p. ej., NYSE TAQ) disponibles en cada instante, pero es cada vez más común que se ofrezca información sobre las k mejores ofertas (p. ej., BME Tick Data ofrece k=5).

El registro electrónico y abierto que recoge todas las ofertas de compra y venta disponibles para un determinado activo en cada instante se denomina "libro de órdenes". Este libro constituye la única fuente de liquidez del mercado, y toda la negociación se desarrolla alrededor del mismo. Es poco habitual poder disponer de datos de todos los niveles del libro. No obstante, estudios empíricos demuestran que en términos de provisión efectiva de liquidez y de formación de precios, con un k relativamente pequeño es suficiente (p. ej., Cao *et al.*, 2008; Brogaard *et al.*, 2019).

Una limitación más significativa es la existencia de profundidad oculta. Muchos mercados financieros a priori transparentes ofrecen la posibilidad de introducir órdenes de volumen oculto (Chakrabarty *et al.*, 2024b), que permiten ocultar al resto de participantes la totalidad o parte del tamaño total de una oferta, por lo que una parte significativa del libro no es visible. La gran mayoría de bases de datos no reportan dicha liquidez oculta. La Figura 4 muestra la distribución del volumen oculto en el libro de órdenes del Nasdaq para días de alta y baja volatilidad. Estos estadísticos se han obtenido de la base de datos Nasdaq HFT (p. ej., Carrion, 2013), que incluye datos del libro de órdenes de 120 activos del Nasdaq entre 2008 y 2010. La Figura 4 muestra que en días de volatilidad alta, la mitad de la profundidad del libro a las mejores cotizaciones está oculta. Este porcentaje cae al 30 % en días de baja volatilidad. El porcentaje de profundidad oculta va disminuyendo a medida que nos alejamos de las mejores cotizaciones.

[Figura 4]

Ejemplos de estudios que utilizan datos de Nivel2 serían Biais *et al.* (1995), que usan datos de *Paris Bourse* para estudiar las dinámicas de la provisión de

liquidez de un libro de órdenes electrónico; Cartea *et al.* (2019), que estudian la relación entre la negociación en alta frecuencia y la calidad del mercado; Næs y Skjeltorp (2006), que muestran que una mayor pendiente (i.e., menor elasticidad) del libro (i.e., menor liquidez) en el *Oslo Stock Exchange* de Noruega indica desacuerdo entre los inversores, intensificando la relación entre volumen y volatilidad; Penalva y Tapia (2021) que estudian la heterogeneidad en la importancia del tamaño del tick y su relación con la competencia entre plataformas de negociación, y Cont *et al.* (2014) y Pascual y Veredas (2010), que aportan evidencia de la capacidad del estado del libro de órdenes para predecir, respectivamente, rentabilidad y volatilidad en el corto plazo.

¿De qué volumen de datos estamos hablando? Tomemos como referencia la base de datos TAQ del NYSE, la base de alta frecuencia más utilizada en la investigación en Microestructura (Hussain *et al.*, 2023). El fichero diario *TAQ Trades*, que recoge un día completo de transacciones (Nivel1), tenía un tamaño medio de 200MB en 2014 una vez comprimido, e incluía 24 millones de entradas. Este mismo fichero en 2024 tiene un tamaño de 2,4GB y tiene 69 millones de entradas. El fichero diario *TAQ Quotes*, que recoge las mejores cotizaciones en el NYSE (Nivel2) a lo largo de una sesión para todos los activos admitidos a negociar en el mercado norteamericano, tenía un tamaño de 6GB una vez comprimido y 550 millones de entradas en 2014. Diez años después, su tamaño ha aumentado a 38GB y contiene 1900 millones de entradas.[13]

(c) Mensajes (datos de "Nivel3"): Engle (2000) llamó *ultra-high-frequency* a la frecuencia en que se registran las transacciones, obviando que otros eventos en la actividad de los mercados financieros ocurren con una frecuencia aun mayor. Este es el caso de los mensajes de negociación, que han acaparado una significativa atención en la última década a raíz de la eclosión de la negociación de alta frecuencia o HFT (O'Hara, 2015). Por mensaje de negociación entendemos cualquier instrucción que un *trader* introduce en una plataforma de negociación electrónica, como puede ser el envío de una nueva orden de negociación, la revisión de las características (precio, tamaño) de una orden almacenada en el libro de órdenes, o su cancelación.

En la Figura 5 mostramos el cociente entre número de mensajes y transacciones para Telefónica S.A. (TEF), uno de los activos tradicionalmente más líquidos de Bolsas y Mercados Españoles (BME), desde julio de 2000 hasta diciembre de 2019. Limitamos el cómputo de mensajes a aquellos que alteran las mejores cotizaciones de compra y venta de TEF en este mercado. Puede observarse cómo desde finales de 2007 el número de mensajes necesario para cerrar una operación se ha incrementado exponencialmente, siendo aproximadamente 1,8 mensajes por transacción en 2000, 2,8 a finales de 2007 y 25 en 2019. Este ejemplo ilustra cómo el crecimiento de la negociación algorítmica, facilitado por la creciente fragmentación de los mercados Europeos que se produce tras la entrada en vigor de la *Markets in Financial Instruments Directive* (MiFID) en

[13] Según Daily TAQ Client Specifications

noviembre de 2007, ha aumentando el volumen de mensajes por segundo enviados a los mercados electrónicos y, por ende, el volumen de las bases de datos de Nivel3.

[Figura 5]

Las bases de datos de Nivel3 proporcionan información mensaje por mensaje. Identifican el tipo de orden ("límite" o "de mercado"), el tipo de mensaje (envío, revisión, cancelación), condiciones especiales ("volumen oculto", "ejecutar o anular", "volumen mínimo"), y el momento del tiempo en que se introduce el mensaje.[14] Con una base de datos de Nivel3 podemos reconstruir en libro de órdenes en su totalidad, tanto la parte mostrada como la oculta. Podemos identificar todas las transacciones realizadas emparejando órdenes entrantes con otras almacenadas en el libro, siguiendo escrupulosamente los criterios de prioridad de precio-tiempo y de volumen mostrado sobre volumen oculto comunes a todas las plataformas de negociación electrónica. Además, dado que conocemos la secuencia en que se han introducido los diferentes mensajes, podemos determinar el signo de cada transacción (compra/venta si la orden entrante toma liquidez del lado bid/ask del libro).

Desafortunadamente, estas bases no suelen estar públicamente disponibles, es decir, no forman parte de la oferta de datos históricos que los académicos pueden adquirir directamente de los mercados o de otros proveedores de datos.[15] Normalmente, los datos de Nivel3 son datos "en propiedad", esto es, que son cedidos por el propietario al equipo de investigación con el único fin de realizar un proyecto de investigación en particular. Como consecuencia, los resultados de estos estudios, siendo altamente valiosos, no pueden ser replicados con facilidad.

Algunos ejemplos de estudios que usan datos de Nivel3 son los siguientes. Brogaard *et al.* (2019) usan datos del *Toronto Stock Exchange* para medir la contribución a la formación de precios de distintos tipos de órdenes, poniendo especial hincapié en las órdenes introducidas por los *traders* de alta frecuencia (HFTs). (Aquilina *et al.*, 2022) usan datos del *London Stock Exchange* para cuantificar la importancia del arbitraje de baja latencia. Su base de datos no sólo contiene los mensajes que alcanzan el libro (exitosos), sino también aquellos que el mercado anula por imposibilidad de ejecutarlos (fallidos), como por ejemplo un mensaje de cancelación de una orden que llega cuando ésta ya ha sido ejecutada. Chakrabarty *et al.* (2024a) utilizan datos del *National Stock Ex-*

[14] Una orden límite especifica precio máximo (mínimo) al que se está dispuesto a comprar (vender), no garantizando ejecución inmediata. Una orden de mercado demanda ejecución inmediata al mejor disponible en el momento en que la orden alcanza el libro. La información sobre volumen oculto está disponible en los L3 de algunos mercado, p.ej. NSE, pero no en otros, p.ej. NASDAQ-ITCH.

[15] Por ejemplo, no hay datos de Nivel3 públicamente disponibles para mercados de EE.UU. Para el caso Europeo es relativamente sencillo encontrar datos de Nivel2, pero no de Nivel3. BME Market Data, por ejemplo, ofrece datos de Nivel1 y Nivel2 solamente

change of India (NSE) para estudiar el uso de órdenes de volumen oculto por los *algorithmic traders* (ATs). Finalmente, Kwan *et al.* (2024) utilizan técnicas de aprendizaje automático (*machine learning*) y datos del *Australian Stock Exchange* para estudiar cómo la formación del precio se ve afectada por el estado del libro de órdenes.

Una de las características distintivas de los datos de alta frecuencia es que los acontecimientos que recogen no se distribuyen uniformemente en el tiempo. Los cambios en las cotizaciones así como las transacciones ocurren como consecuencia del continuado flujo de mensajes, que llega al libro de órdenes a diferentes intensidades para diferentes activos y/o en diferentes momentos de la sesión (p. ej., Easley *et al.*, 1997; Engle y Russell, 1998). En la actualidad, en gran medida debido a la prevalencia de la negociación algorítmica, estas frecuencias se capturan a nivel de microsegundos, milisegundos o nanosegundos, proporcionando una visión extremadamente granular de la actividad del mercado (p. ej., Hasbrouck, 2018, 2021). Un muestreo a intervalos fijos de tiempo de los datos de alta frecuencia supone obviar muchos acontecimientos, especialmente para activos muy negociados y/o en periodos con un flujo de mensajes muy intenso. Ésta es una constante de los datos de alta frecuencia: podemos agregarlos para facilitar su análisis o modelización. Pero toda agregación supone una pérdida de información.

Los datos de alta frecuencia, independientemente de su nivel, suelen contar con carencias comunes. En primer lugar, las bases de datos públicamente disponibles no ofrecen la identidad de los *traders*, debido a que la negociación es, por lo general, anónima. Sí existen bases de datos, normalmente en propiedad, que ofrecen identificadores de tipo de *trader* (p. ej., minoristas, institucionales, algorítmicos etc.).[16]

En segundo lugar, como hemos comentado anteriormente, la negociación en los mercados financieros actuales está muy fragmentada, esto es, repartida entre plataformas alternativas de negociación, algunas transparentes y otras opacas, algunas multilaterales y otras bilaterales. En la Figura 6 mostramos el grado de fragmentación de BME desde 2012 hasta 2022, según datos recogidos de *Battlemap* de Fidessa. Medimos el grado de fragmentación por la cuota de BME sobre el volumen consolidado (en euros) de sus propios activos. En agosto de 2012 la Bolsa de Madrid tenía casi una posición de monopolio en cuanto a la negociación de acciones de empresas españolas, con una cuota de mercado de aproximadamente el 92 %. Tras el levantamiento de las restricciones a la venta en corto en enero de 2013, otros centros de negociación comenzaron a ganar cuota de mercado. En septiembre de 2022, la cuota de mercado de la Bolsa de Madrid se situaba en mínimos históricos, alrededor 61 %.

[16] Ver por ejemplo la base de datos del NSE utilizada por Chakrabarty *et al.* (2024a). En muchos estudios, los investigadores tienen que recurrir a métodos aproximativos para aislar la negociación de algunos participantes (p. ej., Hasbrouck y Saar, 2013; Collin-Dufresne y Fos, 2015; Boehmer *et al.*, 2021)

[Figura 6]

Esta creciente fragmentación añade dificultades adicionales a la hora de analizar los datos de alta frecuencia, especialmente en Europa, donde la falta de una *Consolidated Tape* similar a la de EE. UU. hace muy complicado disponer de datos consolidados completos y fiables de actividad, liquidez y precios para los activos más líquidos negociados en los distintos mercados nacionales, que son los que normalmente experimentan un mayor grado de fragmentación.[17] [18]

Finalmente, modelizar series temporales de datos de alta frecuencia obliga al analista a enfrentarse a características que son propias de estas series, como pueden ser las pronunciadas regularidades intradía, el impacto de los anuncios públicos macroeconómicos y microeconómicos, las particularidades de determinados tramos horarios, como las aperturas, las interrupciones del proceso de negociación por violaciones de límites de precios dinámicos y estáticos (Abad y Pascual, 2010), la persistencia en volatilidad y liquidez, la dificultad de calcular rentabilidades en activos poco negociados, o el ruido en los cambios de precios debido a múltiples fricciones (Stoll, 2002; Hasbrouck, 2007), por citar algunos.

A pesar de todas sus limitaciones, los datos de alta frecuencia ofrecen una precisión inigualable en el análisis de la actividad del mercado y son indispensables para los *traders* y analistas que operan en marcos temporales intradía, para los *traders* algorítmicos, para los académicos que analizan la microestructura del mercado, y para las instituciones que velan por el cumplimiento normativo.

2.2. Datos no estructurados

La era de la digitalización ha generado un incremento sin precedentes de fuentes de información alternativa y, por ende, de los datos no estructurados. Como resultado directo, han surgido un gran número de empresas dedicadas a recolectar, limpiar, analizar e interpretar estos datos y ofrecerlos como producto para apoyar la toma de decisiones de inversión. Según la página web www.alternativedata.org, el número de proveedores de datos alternativos en EE.UU. creció de 150 a finales de 2010 a 445 a finales de 2018. Además, el montante invertido por fondos de inversión en datos alternativos a finales de 2020 rondaba los dos mil millones y el número de empleados especializados en análisis de éstos datos aumentó un 450 % entre 2014 y 2018.

Los analistas financieros desempeñan un papel clave en los mercados financieros. Su responsabilidad principal implica recopilar y analizar información sobre

[17] BME Market Data, por ejemplo, sólo proporciona datos de alta frecuencia correspondientes a la plataforma electrónica del mercado español (*SIBE Smart*).

[18] El Parlamento Europeo ha dado los primeros pasos hacia una *consolidated tape* exclusivamente para transacciones. Para más información, véase aquí.

empresas que cotizan en bolsa y difundir sus conocimientos a los inversores. Chi *et al.* (2023) aportan evidencia de que los analistas financieros han incorporado el uso de datos alternativos a la hora de hacer sus análisis. Además, muestran que cuando su predicciones de beneficios se apoyan en datos alternativos, su precisión mejora. Los intermediarios que suscriben los servicios de estos analistas también cobran más en comisiones de los inversores, sugiriendo que éstos valoran el uso de datos alternativos por parte de los analistas.

La Figura 7 muestra los tipos de datos alternativos más utilizados por fondos de inversión en EE.UU. En la Tabla I aportamos una breve descripción de las diferentes categorías. En general, las fuentes de datos alternativos se pueden clasificar en tres categorías, dependiendo de si son producidas por individuos (p. ej., publicaciones en redes sociales, encuestas, valoraciones de productos, uso de aplicaciones, búsquedas en la red), generadas a través de procesos comerciales o nuevas tecnologías (p. ej., datos de tarjetas de crédito, *web scrapping*), o producidas por sensores (p. ej., satélites, drones, geolocalización).[19]

[Figura 7]

[Tabla I]

Del mismo modo que la industria financiera recurre a los datos alternativos para mejorar sus predicciones y apoyar la toma de decisiones de inversión, los académicos que estudian los mercados financieros recurren cada vez más al uso de estas nuevas fuentes de información, ya sea en forma de texto, imagen, o sonido. En la Tabla II aportamos ejemplos ilustrativos de trabajos académicos que utilizan análisis de texto.

[Tabla II]

El uso de información en redes sociales especializadas (p. ej., StockTwits) o generalistas (p. ej., X/Twitter) es cada vez más común en la investigación en Mercados Financieros. La plataforma StockTwits es una red social en la que los inversores comparten información y opiniones sobre empresas individuales. Dessaint *et al.* (2024) utilizan el lanzamiento de StockTwits como *shock* exógeno para estudiar cómo la creciente disponibilidad de datos alternativos, que ofrecen mayormente información orientada hacia el corto plazo, ha afectado a la calidad y precisión de las predicciones de los analistas financieros, tanto a corto como a largo plazo. Cookson y Niessen-Ruenzi (2020) utilizan más de 18 millones de mensajes de StockTwits entre 2013 y 2014 (de 107 mil usuarios y sobre 9755 activos) para construir un indicador directo de desacuerdo entre inversores basado en el sentimiento (comprador, vendedor) explícitamente expresado por los

[19] Esta clasificación sigue el 2019 Handbook of Alternative Data, de J.P. Morgan

participantes (ver también Giannini *et al.*, 2019). Cookson *et al.* (2020) realizan un análisis lingüístico de los tweets para identificar la orientación política de los participantes en la plataforma y estudiar cómo ésta afecta a su sentimiento (optimista, pesimista) como inversor hacia diferentes activos durante la pandemia del COVID-19. Finalmente, Cookson *et al.* (2022) encuentran que inversores con una determinada creencia es más probable que sigan a otros con creencias similares que opuestas, generando "*echo chambers*" en la plataforma. Las opiniones generadas en estos *clusters* de sentimiento tienen repercusión sobre la actividad en el mercado.

Estos estudios ilustran el valor añadido de los datos no estructurados. Medir divergencias de opinión a partir de datos estructurados es una tarea compleja. La literatura (p. ej. Berkman *et al.*, 2009) recurre a medidas indirectas basadas en el volumen histórico, la volatilidad del beneficio contable o de las rentabilidades, y la dispersión en las predicciones de analistas. Ninguna de estos indicadores recoge explícitamente la opinión de los inversores, incluso en el último caso, sólo se recoge la opinión de los analistas. El uso de datos en redes sociales permite medir el sentir de los inversores directamente, a partir de sus opiniones volcadas en la red. Pero, ¿son estas opiniones realmente informativas? Chen *et al.* (2014) estudian 6500 artículos de opinión publicados la red social orientada hacia el inversor *Seeking Alpha*, así como 180.000 comentarios de los lectores escritos en respuesta a dichos artículos, y encuentran que ambos pueden predecir rendimientos futuros y sorpresas en los anuncios de beneficios.

Uno de los usos más comunes del análisis de texto es el de construir indicadores de sentimiento.[20] En este sentido, el enfoque más común en la literatura es el del "bag-of-words"(BoW) (p. ej., Tetlock, 2007; Jegadeesh y Wu, 2013), en el que a partir de un diccionario de términos financiero-contables, previamente clasificados como términos con connotación positiva o negativa, se mide el tono de un texto en base a un conteo de términos. García *et al.* (2023) combinan tres tipos de texto, transcripciones de telecomunicaciones sobre resultados (*earnings conference calls*) (lenguaje hablado), informes anuales (10-k) (lenguaje legal), y artículos del *Wall Street Journal* (WSJ) (lenguaje periodístico) y técnicas de *natural language processing* (NLP) para generar un diccionario de palabras con connotación positiva o negativa aplicado a finanzas. En lugar de que sean humanos los que codifiquen los términos, aquí es un algoritmo de ML el que evalúa potenciales diccionarios en función de las reacciones de precios alrededor de anuncios de beneficios dentro de una muestra de entrenamiento. Los autores muestran que su diccionario funciona mejor fuera de muestra que los BoW tradicionales.

Chen *et al.* (2023) utilizan Modelos de Lenguaje de Gran Escala (LLM por sus siglas en ingés), como ChatGPT (de *OpenAI*), LLaMA y RoBERTa (de *Meta*) y BERT (de *Google*), para extraer representaciones contextualizadas a partir

[20] Para una revisión en profundidad de esta literatura, véase Gentzkow *et al.* (2019) y Loughran y McDonald (2020).

del texto de noticias para predecir rendimientos. A diferencia de los métodos basados en palabras (BoW), estos modelos capturan tanto la sintaxis como la semántica del texto; se entrenan con grandes conjuntos de datos que abarcan muchas fuentes y temas, miles de millones de parámetros, y miles de millones de ejemplos de texto. Su estimación se realiza una sola vez, por expertos, y luego el modelo estimado se pone a disposición para su distribución (software de código abierto) para ser utilizado por investigadores no especializados, quienes solo necesitan alimentar el modelo con el documento de interés. Utilizando datos de 16 países y textos en 13 idiomas diferentes, estos autores muestran que estos modelos tienen mayor poder predictivo sobre los movimientos del mercado que modelos tradicionales, tanto de análisis técnico como otros modelos de NLP.

Medir sentimiento no es la única utilidad del análisis lingüístico. Por poner algunos ejemplos recientes, Fedyk (2023) estudia el efecto que el posicionamiento de una noticia en los terminales de *Bloomberg* puede tener sobre la rapidez con que la información contenida en dicha noticia se incorpora a los precios. Manela y Moreira (2017) construyen un indicador de incertidumbre basado en noticias de portada del WSJ. Akey *et al.* (2022) miden el contenido informativo en comunicados de prensa sobre anuncios de beneficios a los que piratas informáticos (*hackers*) tuvieron acceso ilegalmente antes de su publicación. Sautner *et al.* (2023) utilizan transcripciones de teleconferencias sobre resultados para medir la percepción que los participantes tienen sobre la exposición de la empresa a riesgos asociados a diferentes facetas del cambio climático. Los indicadores resultantes contienen información relevante sobre el valor de las acciones, además de poder explicativo sobre las políticas medioambientales emprendidas por la empresa. Hillert *et al.* (2014) construyen un indicador de divergencia de opinión entre periodistas basado en artículos publicados en *The New York Times*, *The USA Today*, WSJ, y *The Washington Post*. Davis *et al.* (2020) miden factores de riesgo a nivel de empresa utilizando informes anuales (10-K *filings*) registrados en la SEC antes del inicio de la pandemia del COVID-19. Con ellos, explican la diversidad de reacciones en el precio de las acciones en respuesta a la pandemia.

En la Tabla III aportamos ejemplos ilustrativos de trabajos que utilizan datos no estructurados que no son en forma de texto. El uso de imágenes como input, facilitado por el desarrollo de algoritmos de *deep learning*, como los *convolutional neural networks* (CNN) (ver Sección 3), es un fenómeno más reciente que el análisis de texto. Los limitados ejemplos existentes, sin embargo, ilustran el potencial de estas técnicas aplicadas al estudio de los mercados financieros. Así, Deng *et al.* (2023) muestran teórica y empíricamente que incluir contenido gráfico en la versión "amigable" de informes anuales tiene efectos positivos en los rendimientos anormales de las empresas en los siguientes 3-6 meses. Los autores concluyen que las empresas incluyen el contenido visual para captar la atención del inversor y, al mismo tiempo, comunicar información relevante a los inversores fundamentales.

En una aplicación de análisis técnico, Jiang *et al.* (2023) identifican en gráfi-

cos pixelados de series históricas de precios diarios (apertura, máximo, mínimo, cierre) los patrones subyacentes con mayor poder de predicción. Estos patrones son en ocasiones tan sutiles que escapan a otros métodos más tradicionales y suficientemente complejos como para que un ser humano no los pueda anticipar. Obaid *et al.* (2022) construyen un índice de sentimiento diario a nivel global del mercado ("Foto Pesimismo") basándose en una clasificación de fotos publicadas en el WSJ. Este índice considera diferentes características de las fotos, como el color, las expresiones faciales, los objetos que aparecen etc. En linea con modelos de Finanzas del Comportamiento, el foto pesimismo predice caídas de precios y aumentos de volumen. Finalmente, Gerker y Painter (2022) utilizan imágenes satelitales de zonas de aparcamiento de empresas minoristas para medir el performance de éstas en localizaciones concretas. Con estos datos documentan que los analistas financieros tienden a basar sus recomendaciones en señales próximas a su localización geográfica cuando no existe información global sobre el performance de la empresa.

[Tabla III]

Alejándonos de los mercados financieros bursátiles, encontramos aplicaciones adicionales del análisis de imágenes. Aubry *et al.* (2023) estudian el caso de las subastas de obras de arte. Basándose en características visuales y no visuales (características del artista, de la obra, y de la casa de subastas) de 1,2 millones de obras objeto de subastación, estos autores utilizan un modelo de CNN para realizar predicciones de precios de asignación durante la subasta, y aportan evidencia de sesgos sistemáticos en las casas de subastas hacia la obra de ciertos artistas. Glaeser *et al.* (2018) estudian, para el caso del mercado inmobiliario en Boston (EE.UU.), hasta qué punto la apariencia externa de una casa o de las casas vecinas influye sobre su precio de venta. Para ello utilizan imágenes de las mismas viviendas a lo largo del tiempo, proporcionadas por *Google Street View*.

Los anteriores estudios muestran que hay información valiosa que subyace en textos e imágenes. Del mismo modo, los investigadores pueden extraer información útil a partir ficheros de audio o vídeo. Edmans *et al.* (2022) introducen una medida de sentimiento a nivel nacional basada en el positivismo de las canciones que las personas eligen escuchar a través de servicios de *streaming* como *Spotify* o *Amazon Music*. El sentimiento musical capta cambios de ánimo y afecta a la rentabilidad en los mercados de valores, tanto de renta variable como de renta fija. Además satisface las tres características del BD identificadas por Goldstein *et al.* (2021): gran tamaño (ya que agrega el comportamiento de escucha diario de todos los clientes de Spotify dentro de un país), alta dimensión (dado que una canción tiene múltiples características que contribuyen a su medida de valencia - carácter emocional positivo o negativo de la canción), y complejidad. A partir de los audios de las teleconferencias sobre resultados, Mayew y Venkatachalam (2012) proporcionan evidencia de que el tono vocal y el estado de ánimo de los directivos contiene información útil sobre los fundamentos de una empresa.

Finalmente, Huang *et al.* (2023) estudian el efecto de las primeras impresiones de emprendedores en las decisiones de inversores privados ("*angel investors*"). Utilizando imágenes extraídas de vídeos de presentaciones en los programas televisivos *Shark Tank* y *Startup Batterfield*, miden las habilidades generales, el encanto, y la capacidad de gestión del emprendedor.

En la era de la digitalización, los académicos disponen de una variedad de fuentes de datos alternativos sin precedentes, que poco a poco están incorporándose a la investigación en Mercados Financieros. Por ejemplo, Gibbons *et al.* (2021) utilizan la frecuencia de uso de la base de datos *Electronic Data Gathering, Analysis, and Retrieval* (EDGAR) para medir adquisición de información pública por parte de los analistas financieros y cómo ésta afecta a la calidad de sus recomendaciones. Da *et al.* (2015) construyen una medida de sentimiento a partir de las búsquedas en *Google* de millones de usuarios. Su índice FEARS (*Financial and Economic Attitudes Revealed by Search*) predice caídas de precio, volatilidad, y transferencias de fondos del mercado de renta variable al mercado de renta fija. Di Maggio *et al.* (2020) utilizan los registros de las cajas registradoras de los supermercados para estudiar las decisiones financieras de los hogares. Finalmente, Froot *et al.* (2017) se apoyan en medidas de la actividad del consumidor en tiempo real construidas por *MKT-Mediastats* a partir de múltiples fuentes, incluidos millones de dispositivos electrónicos de consumidores (teléfonos y tabletas), para predecir el crecimiento de las ventas, los ingresos, y el componente no esperado de las ganancias de las empresas.

Una parte integral del proceso de recogida de datos, especialmente cuando el volumen de éstos es grande, es la codificación de la información. Antes de empezar a procesar la información ésta tiene que estar en un formato susceptible de ser procesado. La creciente digitalización de la sociedad facilita la recogida de estos datos a todos los niveles. Por ejemplo, cada vez es más común el uso de aplicaciones, tanto para móviles como para ordenador, en la toma de decisiones de inversión. A través de estas plataformas de inversión en red se recogen grandes volúmenes de datos, tanto de las decisiones en sí (cuánto y en qué invertimos), como de las circunstancias que rodean estas decisiones (la hora exacta, la información que miramos en la aplicación antes de invertir, las noticias que leemos, etc.). La contrapartida es que este proceso genera una variedad y volumen de información tal que su uso supone un reto para el investigador.

No obstante, la investigación financiera avanza y va incorporando gradualmente estas nuevas fuentes de información. Así, para estudiar las decisiones de los pequeños inversores Eaton *et al.* (2022), Bartlett *et al.* (2024), Welch (2022), y Barber *et al.* (2022) utilizan datos públicos de la plataforma *Robinhood* que contienen estadísticas de los activos en los que sus usuarios invierten. La plataforma *bitly* recoge y digitaliza información con vínculos a diferentes páginas web y artículos en la red. Esta inmensa información la utilizan Benamar *et al.* (2021) para estudiar la demanda de información y la incertidumbre alrededor de anuncios relacionados con los bonos del Tesoro en EE.UU. y su impacto sobre

los precios. Gargano y Rossi (2024) proporcionan evidencia de que las aplicaciones de *FinTech* pueden mejorar las decisiones de consumo y ahorro de las personas al ayudarlas a establecer metas de ahorro. Finalmente, Carlin *et al.* (2023) utilizan el lanzamiento de una aplicación para móvil que agrega información financiera del usuario (cuentas de ahorro, tarjetas etc.) para estudiar cómo facilitar el acceso a la información ayuda al consumidor a tomar mejores decisiones.

3. Métodos

En paralelo al crecimiento del tamaño, dimensión, y complejidad de los datos, hemos experimentado también un amplio desarrollo en la accesibilidad, variedad, y potencia de los recursos físicos y métodos para tratar estos datos. Por un lado la investigación ha aumentado la capacidad y velocidad de los recursos informáticos (ordenadores, redes, almacenaje, etc.) tanto localmente como de manera virtual (p. ej., AWS, Google Cloud, Azure). Por otro lado, las bibliotecas de métodos han aumentado notablemente en su variedad, accesibilidad, y capacidad de cálculo, empezando por lenguajes de programación (Python) y de análisis de datos (R). Estos lenguajes gratuitos y de código abierto dan acceso a grandes bibliotecas de métodos para el análisis y procesamiento de datos que permiten gestionar y aprovechar su riqueza.

El abanico de métodos existente es muy rico y de difícil clasificación. Por un lado se puede intentar clasificar en base a diferencias en su estructura. Podríamos diferenciar entre métodos lineales y no lineales, o entre métodos para espacios discretos versus contínuos. También podríamos diferenciar entre métodos con una función objetivo global, versus métodos con funciones objetivo dinámicas. Pero con el tiempo los métodos se van hibridando, y combinando, de tal manera que las combinaciones de variantes no son fácilmente distinguibles. Por otro lado se puede intentar clasificar por su uso. Unos métodos son más apropiados para clasificar información, otros para resumirla, mientras que hay métodos diseñados para la predicción. Pero también esta estructura no es válida ya que hay métodos que se pueden adaptar para diferentes usos. Nuestro texto utiliza una estructura que hila diferentes métodos aprovechando ámbitos de aplicación (extracción, clasificación, modelización, predicción, y optimización) y similitudes de diseño, mientras repasamos sus aplicaciones a problemas financieros, como pueden ser los índices de sentimiento o confianza, el estudio y predicción de rentabilidades, el análisis de inversiones, y la gestión de carteras.

Empezamos por métodos inicialmente diseñados para la extracción y clasificación de información, y mostramos aplicaciones a los Mercados Financieros. Éstas últimas nos llevan a hablar de las redes neuronales, que aunque tienen usos en todos los ámbitos de ML, las vemos en el contexto de clasificación y modelización de datos. De ahí pasamos a discutir métodos de predicción, cen-

trándonos en el problema de *overfitting* y la teoría y métodos de *shrinkage* que engloban una parte importante de las aplicaciones en Mercados Financieros, y enespecial en valoración de activos y fondos. Concluimos repasando los métodos que combinan aprendizaje y optimización agrupados bajo el paraguas de *Reinforcement Learning*. En la Tabla IV se resumen las principales metodologías y un ejemplo representativo.

[Tabla IV]

3.1. Representation Learning

Uno de los retos más significativos que genera la amplia variedad de fuentes de datos disponibles en la era digital, discutidas en la Sección 2, es el de la alta dimensión, esto es, el gran número de variables potencialmente relevantes con las que tiene que lidiar el investigador. Éste necesita identificar las principales fuentes de información en los datos, para así poder reducir el número de variables potenciales a uno manejable (para el investigador), lo que se denomina *representation learning*. Dos grupos de técnicas frecuentemente utilizados para obtener esta reducción son el filtrado (*filter*) y la envoltura (*wrapper*).

Los métodos de filtrado buscan identificar las variables que recogen la mayor cantidad de información posible, normalmente sin hacer referencia al problema específico que queremos resolver, utilizando diferentes medidas de información (entropía, información de Fisher, correlación con la variable de interés, etc.). Para ello, se aplican métodos desarrollados en el área del álgebra lineal como PCA (*Principal Components Analysis*), PCR (*Principal Components Regression*), y PLS (*Partial Least Squares*). Estos métodos utilizan factorizaciones de matrices para identificar las combinaciones más informativas de las variables.

Falta la más importante : eliminar atributos con varianza cercana a 0

Las técnicas de PCA y PCR se han utilizado para construir índices de opinión (Baker y Wurgler, 2006, 2007) e índices de riesgo (Jurado *et al.*, 2015), así como para estimar factores de riesgo (Ludvigson y Ng, 2007, 2010). En PLS, en lugar de extraer información de manera genérica, como en los métodos anteriores, el objetivo es identificar la información más relevante para predecir una variable de interés. Por ejemplo, Chatelais *et al.* (2023) extrae información de precios de activos por industria para predecir la actividad económica. Light *et al.* (2017) utiliza PLS para predecir rentabilidades en base a las características de las empresas. Chen *et al.* (2022), Huang *et al.* (2015), y Kelly y Pruitt (2013) combinan la información de varios índices de opinión para predecir los movimientos del mercado y la prima de riesgo, mientras que Kelly *et al.* (2023) usa información de carteras para predecir rentabilidades. Existen múltiples variantes de estas técnicas, como *Scaled* PCA, propuesta por Huang *et al.* (2022), así como extensiones a problemas con procesos estocásticos. Por ejemplo, De Spiegeleer *et al.* (2018) aplica métodos de aprendizaje basados en regresiones con procesos

Gaussianos para valorar opciones y su cobertura.

Los métodos de envoltura evalúan diferentes combinaciones de variables de manera secuencial, y dentro del contexto concreto del modelo de análisis que se va a utilizar, para identificar aquellas que generan un mejor resultado. Este método lo utilizan algoritmos como *Support Vector Machines* (SVM), o los árboles de decisión (*regression trees*, *random forest*). Los SVM están basados en la geometría, ya que buscan dividir el espacio de datos mediante hiperplanos. Cada hiperplano separa el espacio en dos grupos de tal manera que la distancia entre los dos grupos es máxima. Este tipo de algoritmo se denomina de clasificación lineal, aunque si aplicamos el conocido como *kernel trick*, que consiste en transformar los datos de una manera no lineal, se puede conseguir una separación entre grupos más compleja y efectiva. Por contra, los árboles de decisión utilizan estructuras gráficas para extraer información y/o clasificar los datos. También separan los datos, aunque en este caso lo hacen de variable en variable, dividiendo la muestra en dos grupos de la manera más efectiva posible. Los árboles de decisión son la base para otros algoritmos más sofisticados como los que utilizan las técnicas de *boosted gradient regression trees* y *random forest*.

"dividiendo la muestra en dos grupos": los métodos de árbol no solo dividen en 2 grupos. Yo lo dejaría como "dividen en grupos". Por ej el algoritmo C4.5

"boosted gradient regression trees" mejor como "gradient boosted trees". Por ej XGBoost no solo se usa en regresión sino en clasificación.

Tanto las SVM como las técnicas gráficas basadas en árboles de decisión se usan para reducir la dimensión de los datos por agrupación, y se añaden a técnicas ya existentes utilizadas para clasificar datos de manera óptima, como los modelos *probit/logit*, de gran utilidad en múltiples aplicaciones en Finanzas. Por ejemplo DeMiguel *et al.* (2023) utiliza estas técnicas para clasificar fondos de inversión en base a sus características y predecir rentabilidades; Mittnik *et al.* (2015) para identificar factores macroeconómicos y financieros que afectan a la volatilidad de los activos; Tobeck y Hronec (2021) para agregar las carteras de anomalías en un solo factor de *mispricing* en la valoración de activos; Lin *et al.* (2023) usan árboles adaptados a datos de panel para construir una mejor frontera eficiente y valorar activos, y Griffin *et al.* (2023) para analizar los márgenes de los dealers de bonos municipales. Luss y D'Aspremont (2015) utiliza texto de noticias para predecir grandes movimientos en precios de acciones utilizando SVM con múltiples *kernels*. Easley *et al.* (2021a) utilizan *random forest* para mostrar que, en el contexto de los complejos mercados financieros actuales, medidas estándar en microestructura predicen variables de interés para los participantes en el mercado.

3.2. Redes Neuronales

En el contexto de la extracción de información, además de los métodos de aprendizaje ya vistos, encontramos métodos basados en redes neuronales, que han evolucionado notablemente en los últimos años hasta llegar a los modelos de redes de aprendizaje profundo que hay detrás de ChatGPT, Bing, Gemini, o Anthropic's Claude, y que se asocian popularmente con la IA.

No necesariamente las neuronas deben tener todas sesgo. Una de las funciones de activación más comúnes es la ReLU: “max(0, x)”. De hecho en Kaniel et al. (2023) usan la función ReLU (página 15).

Las redes neuronales se basan en aplicar transformaciones a los datos, por medio de “neuronas”. Cada neurona almacena los pesos en un vector w con los que se ponderan las entradas, y una función de activación para transformar la salida. Si tomamos una observación de nuestros datos, vector x, (formada por la combinación de n variables) una neurona lo que hace es transformar esta observación a una única señal s_i = f (w^T · x_i + b). Con esta transformación lo que se busca es extraer información de una manera no lineal. [El ejemplo sería mejor si fuese]:
Por ejemplo, imaginemos un sistema que debe detectar si mientras conducimos vamos a tener un choque con el coche de delante. La señal a generar sería 0 (no habrá choque) o 1 (sí habrá choque). Las variables serían la potencia del motor, la fuerza de frenado, los años de experiencia, el tipo de asfalto, etc. (nuestro vector x). Definir explícitamente (programando reglas) todas las interacciones no lineales es imposible, por esto las redes neuronales son ideales para definir las relaciones, al capturar en su vector de pesos w las relaciones entre variables y señal.

Las redes neuronales se basan en aplicar transformaciones no-lineales a los datos, por medio de “neuronas”. Si tomamos una observación de nuestros datos, x, una neurona lo que hace es ajustar esta observación de manera lineal (la multiplica por unos pesos w y le añade una cantidad b o “sesgo”), y al resultado del ajuste le aplica una función no-lineal, f, para obtener un resultado final, $s = f(w\,x + b)$. La neurona convierte cada valor de nuestros datos x_i en la señal correspondiente s_i.[21] Con esta transformación lo que se busca es extraer información de una manera no lineal. Por ejemplo, supongamos que queremos determinar cuánto azúcar añadir cuando estamos haciendo helado. Todos tenemos un gusto diferente pero cada uno tenemos un punto de azúcar. Más azúcar y puede estar demasiado dulce, poco azúcar y está soso. Para intentar predecir cuán dulce hacer el helado, necesitamos aplicar una función no-lineal sobre la cantidad de azúcar que lleva. Lo mismo pasa con la combinación de leche y azúcar en el café. La interacción entre estas variables y su efecto requiere una modelización no-lineal que se puede capturar con estas neuronas. Estos métodos permiten lidiar con los problemas que surgen de la alta dimensión en el BD.

La composición de neuronas crea una red neuronal

“Si los bloques los organizamos” : Si la red la organizamos en capas apiladas, obtenemos un DNN

La red neuronal crea bloques de neuronas, donde cada dato se convierte en varias señales, que extraen información de manera no-lineal, y éstas señales a su vez se combinan en un resultado agregado. Si los bloques los organizamos por capas, de tal manera que la red aplica un bloque a los datos, y después aplica otro bloque a los resultados del primer bloque, obtenemos una red neuronal profunda (*Deep Neural Network*, DNN). La red neuronal, al final, actúa reconvirtiendo los datos. Esta reconfiguración de los datos puede resultar en algo más complejo o más sencillo, y está determinado por la estructura de la red y sus capas. Entre estas estructuras encontramos redes neuronales recursivas (*Recursive Neural Nets*), redes neuronales convolucionales (*Convolutional Neural Nets*), *Autoencoders*, etc.

Las redes neuronales se han utilizado en la investigación en Mercados Financieros para múltiples propósitos. Una contribución temprana la encontramos en Brown *et al.* (1998) que utiliza estas técnicas para codificar las editoriales de W.P. Hamilton en el WSJ, que intentaban predecir movimientos en el mercado. Establecen que las editoriales tenían capacidad de predicción y ésta era esencialmente debida al impulso (*momentum*) en los precios. Kaniel *et al.* (2023) utiliza redes neuronales para identificar qué características de los fondos de inversión ayuda a predecir su rentabilidad.

Las redes neuronales también se utilizan para generar representaciones de espacios muy complejos con los que trabajar en diferentes aplicaciones. En el contexto de la modelización de la volatilidad Wiese *et al.* (2020) utilizan un tipo de redes neuronales llamadas redes antagonistas generativas (*Generative Adversarial Networks*) (GAN), combinadas con redes CNN que (como ya vimos en la Sección 2) se suelen utilizar en el procesado de imágenes, para generar modelos

[21] El lector interesado encontrará un tratamiento de las redes neuronales más extenso, preciso, y detallado en Murphy (2022).

de procesos de precios que incorporen dependencias entre series a largo plazo, como por ejemplo, *clusters* en volatilidad. Mientras, Horvath *et al.* (2021) utilizan redes neuronales para calibrar la superficie de volatilidad (*volatility surface*) implícita de las opciones en su totalidad y de una manera eficiente. Brogaard y Zareei (2023) utilizan técnicas evolutivas (*evolutionary genetic algorithms*) para encontrar reglas de análisis técnico que sean beneficiosas, en el sentido de que puedan predecir cambios en los precios. Encuentran que dichas estrategias existen, pero que su efectividad decrece en el tiempo, sugiriendo mejoras en la eficiencia del mercado. Bekiros (2013) usan redes neuronales y Makinen *et al.* (2019) y Tashiro *et al.* (2019) usan CNN para predecir precios.

Un ejemplo de espacio complejo sería el de modelizar las dinámicas del libro de órdenes (ver Sección 2.1). Siguiendo el planteamiento en Bekiros (2013) y añadiendo mucho más detalle sobre el estado del libro de órdenes Kolm *et al.* (2023) aplica diferentes estructuras neuronales, codifica la información en el libro de órdenes, y la utiliza para anticipar movimientos de precios a muy alta frecuencia. Sirignano (2019) utilizan CNN para obtener un modelo de predicción de movimientos de precios de baja dimensión que incluye la información del libro de órdenes y que captura la dinámica de éstos precios en las colas, donde los modelos más sencillos, como los lineales, suelen ser poco efectivos. Kercheval y Zhang (2015) también modelizan la dinámica en el libro de órdenes, que ayuda a predecir cambios en el punto medio y transacciones.

La arquitectura que subyace a un chatbot es un modelo formado por 2 submodelos. El primero es un modelo de embedding, que aprende representaciones vectoriales comprimidas (algo. skip-gram) de un corpus (por ej word2vec o seq2seq) y un segundo modelo, transformer (ya sea encoder-only (BERT), decoder-only (chatGPT) o encoder-decoder.

Al hablar de redes neuronales, no podemos obviar los codificadores (*encoders*), las redes neuronales que subyacen a los *chatbots* más famosos. Los codificadores son estructuras de redes neuronales. A grandes rasgos, el *chatbot* descompone una frase de entrada, por ejemplo, en español, en su código interno cuantitativo ("representación latente"), predice cuál debería ser la frase siguiente en su código interno y la recodifica al español. De tal manera que si interpreta la frase inicial como una pregunta, la frase que predice es la respuesta, lo que genera la ilusión de que responde. Por ello, la lingüista Emily Bender (*University of Washington*) describe estos *chatbots* como loros estocásticos. En la investigación en Mercados Financierios encontramos aplicaciones como Kolm *et al.* (2023), mencionado anteriormente, que utiliza estos codificadores, o auto-codificadores (*autoencoders*), para predecir precios futuros.

NOTA: un encoder y un transformer-encoder NO TIENEN NADA QUE VER. Sus objetivos y arquitecturas son distintas!!!

3.3. Regularización

Como hemos visto, el aprendizaje automático va más allá de la extracción de información, en la medida que ésta se hace de tal manera que una simple clasificación se convierte en una herramienta de predicción. Los métodos que mejor ilustran este aspecto del aprendizaje automático son las regresiones lineales generalizadas (GLR) combinadas con técnicas de regularización. Estas técnicas generalizan la clásica regresión lineal imponiendo penalizaciones al número de

Yo hubiese puesto las técnicas de regularización antes que las de redes neuronales

parámetros con el fin de obtener modelos más sencillos y robustos. La robustez es especialmente importante cuando estamos usando bases de datos con gran cantidad de variables, en cuyo caso sabemos que la regresión lineal tradicional no produce buenas estimaciones o incluso, si el número de variables es superior al de observaciones, ni siquiera se puede utilizar.

Las técnicas de regularización (*LASSO*, *ridge regression*, y *ElasticNet* entre otras) penalizan la estimación lineal de diferentes maneras. Estos métodos están fundamentados en la teoría de matrices aleatorias (véase Ledoit y Wolf, 2004, 2017, 2020, entre otros). Esta sólida base teórica permite integrar avances en aprendizaje automático dentro de modelos de Finanzas teóricos. Por ejemplo, Kelly *et al.* (2024) analizan el *trade-off* entre precisión en las predicciones y la varianza en el error de estimación asociado a modelos "complejos" (i.e., el número de parámetros excede el número de observaciones) comparado con modelos "simples" (pocos parámetros). Teóricamente demuestran que los modelos simples subestiman severamente la predictibilidad de las rentabilidades en comparación con los modelos complejos. Empíricamente, el uso de *ridge regression* mejora el ratio de Sharpe de una cartera de anticipación (*timing portfolio*) cuando el número de variables excede el número de observaciones. Sus resultados teóricos y empíricos proporcionan una justificación para modelizar las rentabilidades esperadas mediante ML. Además Dello Preite *et al.* (2024) incluye un análisis teórico de un modelo de valoración por factores, en el que el método de *ridge regression* aparece como estimador óptimo del modelo de valoración.

Lasso aplica la regularización L1, Ridge la regularización L2 y ElasticNet aplica alpha*L1 + (1-alpha)*L2

L1 y L2 son distintas, pero Elastic es la combinación de ambos

Esto es algo que hubiese hablado al principio del punto 3.
Cuanto más sencillos los modelos (se hacen más suposiciones) mejor es la varianza pero peor el sesgo. Por ej un modelo Ridge asume relación lineal.
Cuanto más complejo el modelo peor es la varianza pero mejor el sesgo. Por ej los modelos de árbol

Pero es en aplicaciones empíricas donde encontramos habitualmente este tipo de modelos. Por ejemplo, en la literatura sobre identificación de factores en modelos estadísticos de valoración de activos, el llamado problema del "zoo de factores", que está generando una literatura demasiado amplia para resumirla en este trabajo.[22], DeMiguel *et al.* (2023) utiliza métodos de aprendizaje automático para estimar el rendimiento de fondos de inversión, entre los que están *ElasticNet* y *random forest*. También hay aplicaciones en otros ámbitos de las Finanzas, como en la construcción de carteras media-varianza robustas Ho *et al.* (2015), o en la predicción de precios en base a datos contables (Akramov *et al.*, 2021).

En el contexto de predicción de rentabilidades esperadas, Gu *et al.* (2020b) hace una comparativa entre diferentes métodos.[23] Su punto de partida es la regresión lineal clásica, y van introduciendo cambios. Primero consideran métodos que modifican el objetivo tradicional de minimizar errores cuadrados, ajustando la función objetivo para reducir el impacto de observaciones en la colas, y así obtener resultados más robustos. De ahí pasan a utilizar los métodos que acabamos de ver, de regresiones con penalizaciones (*Ridge*, *LASSO*, *ElasticNet*). También utilizan PCR y PLS, métodos de regresión generalizados, árboles de regresión, *random forest*, y redes neuronales. De la comparativa extraen que los

[22] El lector interesado, puede leer Bryzgalova *et al.* (2023).
[23] Y explican los métodos con más detalle, para el lector interesado.

métodos no lineales nos ayudan a entender mejor el proceso de formación de precios. De entre éstos, los árboles y las redes neuronales son los métodos que mejor predicen, aunque comparando redes neuronales con pocas capas con las que tienen muchas (*Deep Learning*), la primera es mejor. En general, la mejora con el uso de ML es mayor para activos grandes y muy líquidos. Y todos los métodos encuentran las mismas variables predictivas: *momentum* y reversión de precios, así como la liquidez, volatilidad, y ratios de valoración de los activos.

3.4. Reinforcement Learning

Un grupo de métodos de ML desarrollados recientemente es el de aprendizaje por refuerzo (*Reinforcement Learning*) (RL). Estos métodos combinan las técnicas de aprendizaje vistas anteriormente con funciones objetivo, y aprenden gradualmente, interactuando con un entorno en el que toman decisiones y que les proporciona retroalimentación en forma de recompensas y penalizaciones. Estos métodos son muy utilizados, sobre todo en la literatura de Matemáticas Financieras, como métodos eficientes de optimización para resolver un extenso elenco de problemas en entornos complejos. Por ejemplo, Buehler *et al.* (2019) utilizan métodos basados en aprendizaje profundo combinados con métodos de refuerzo para optimizar la construcción de carteras de derivados. Cartea *et al.* (2022a) construyen estrategias óptimas de negociación entre acciones y tipos de cambio, utilizando firmas temporales (*signatures*) para codificar la dinámica de precios combinadas con RL. Kim y Kim (2020) utiliza ML para minimizar el *tracking error* de una cartera sobre un índice. Wang y Zhou (2020) resuelve un problema de optimización de carteras media-varianza en tiempo continuo con RL. Con ello, consigue identificar la estrategia de retroalimentación óptima. Estos métodos también sirven para modelizar escenarios interesantes. Por ejemplo, Cartea *et al.* (2022b) utilizan ML para modelizar cómo varios algoritmos proveedores de liquidez compiten entre sí.

Los métodos de aprendizaje automático aquí mencionados son una parte representativa de la riqueza de métodos disponibles. Entre los monográficos que tratan estos métodos, en Murphy (2022) encontramos una relación más extensa y detallada de métodos de aprendizaje automático, que incluyen modelos Bayesianos, modelos de grafos probabilísticos, y un largo etcétera.

4. Conclusiones

En los últimos años hemos sido testigos de la irrupción de la Inteligencia Artificial (IA) y del *Big Data* (BD) en diferentes ámbitos de la sociedad. La Ciencia no es una excepción. La IA proporciona nuevas fuentes de datos y nuevas técnicas para trabajar con dichos datos, por lo general de gran volumen, alta

dimensión, y complejidad. En muchos casos, estos datos vienen en formatos poco habituales como texto, imagen, sonido, vídeo, etc., aumentando las fuentes de información útil a disposición del investigador.

En este trabajo hemos aportado numerosos ejemplos que ilustran cómo, en la última década, la IA ha contribuido significativamente a la investigación en Mercados Financieros facilitando mediciones más precisas (o simplemente alternativas) de fenómenos objeto de estudio, identificando relaciones sutiles y complejas entre variables que escapan de la observación con métodos tradicionales, o derivando formas funcionales para describir procesos que van más allá del tradicional modelo lineal y que son difíciles de anticipar en base a la teoría existente.

Disponer de más fuentes de información y de más herramientas para el análisis, y a un coste que, cabe esperar, sea cada vez menor, debería aumentar la capacidad y la precisión con que los académicos podemos contrastar teorías y, por tanto, avanzar en nuestro entendimiento del funcionamiento de los Mercados Financieros. No obstante, este proceso no está exento de potenciales dificultades.

(a) Sesgo hacia el corto plazo: Fundamentándose en una revisión de trabajos empíricos, Dessaint *et al.* (2024) argumentan que la información contenida en las fuentes de datos alternativas, especialmente las no estructuradas, podrían estar orientadas hacia el corto plazo. Estos autores muestran, teórica y empíricamente, cómo un creciente uso de estas fuentes alternativas de información afecta positivamente a la calidad y precisión de las predicciones de los analistas financieros a corto plazo, pero negativamente a las predicciones a largo plazo. Este sesgo podría extenderse también al ámbito de la investigación, limitando la capacidad del investigador orientado a contrastar relaciones de largo plazo, como en el ámbito de la valoración de activos.

(b) Información y ruido: Mayor volumen (de datos) no necesariamente implica mayor calidad y mejor información. Las fuentes de datos alternativas no son ajenas al tradicional problema de separar información de ruido, incluso puede que en ciertos foros, como las redes sociales, el ratio de ruido sobre información sea anormalmente alto. Dugast y Foucault (2018) proponen un marco teórico en el que *traders* negociando rápidamente en base a información bruta, no contrastada, inyectan ruido en precios. En su modelo, procesar la información filtra el ruido, pero requiere de cierto tiempo. Si el coste de la información bruta disminuye, también lo hace la demanda de información procesada y, por ende, la eficiencia en precios a largo plazo. Nuevamente, este argumento podría trasladarse a la investigación. Un procesamiento y filtrado poco cuidadoso de la información disponible en las redes sociales y otras fuentes de datos alternativos no contrastadas podría añadir mucho ruido a la hora de construir variables de interés, realizar predicciones, identificar relaciones etc. Por ello sería importante establecer criterios mínimos y uniformes de calidad de la información en los datos antes de aceptarlos como válidos para la investigación.

(c) Impacto desigual: Mientras que las técnicas de *machine learning* (ML) han sido ya integradas dentro de la "caja de herramientas" propia de ciertas áreas de la investigación en Mercados Financieros, como en el caso de las Finanzas del Comportamiento, la Valoración de Activos, el Análisis Técnico, o el estudio de instrumentos de inversión colectiva, en otras áreas el impacto ha sido menos significativo. Este es el caso de la Microestructura, un área que, por lo general, genera trabajos de investigación que involucran un mayor volumen de datos (de alta frecuencia).[24]

Una posible causa es que las técnicas de ML están orientadas mayormente hacia la predicción, mientras que la Microestructura pone más hincapié en la inferencia. También es posible que el tipo de fenómenos que centran la atención de la Microestructura, como la evaluación de cambios de regulación y organizativos en los mercados, la formación del precio a altas frecuencias, o el estudio de los determinantes de la liquidez de un activo, puedan explicarse de manera más que satisfactoria mediante el uso de técnicas de análisis estadístico y econométrico tradicionales aplicadas a los datos (estructurados) de alta frecuencia, con lo que el valor añadido del ML puede que sea menor. Por ejemplo, Easley *et al.* (2021b) muestran que a la hora de hacer predicciones de precios y liquidez a frecuencias altas utilizando medidas de Microestructura, una simple regresión logística funciona tan bien como complejos modelos de ML.

Finalmente, el aprendizaje automático está orientado hacia la comprensión del comportamiento humano. ¿Qué sucede, sin embargo, cuando la negociación en los mercados financieros no es cosa de humanos, sino de máquinas? Gran parte del volumen negociado y del flujo de órdenes en los mercados financieros modernos está generado por algoritmos. Los ATs dominan la provisión de liquidez y la formación de precios a altas frecuencias. Existe, por lo tanto, una línea de investigación prometedora en la aplicación del ML para identificar y comprender el comportamiento de dichos operadores de alta frecuencia. Por ejemplo Cartea *et al.* (2022c) utilizan técnicas de aproximación estocástica sobre algoritmos de ML para describir teóricamente como se comportan los algoritmos de los operadores de alta frecuencia cuando compiten entre sí, y encuentran que pueden aprender a coludir, entre otros comportamientos competitivos y anticompetitivos. Además, hay diversas líneas de investigación en Microestructura que podrían beneficiarse de la aplicación de técnicas de ML, como por ejemplo, el diseño de *circuit breakers* basado en indicadores anticipados de toxicidad o fragilidad, medidas a partir de técnicas de ML; la identificación de la negociación de tipos particulares de traders (minoristas, HFTs, informados, *insiders*, institucionales etc.), o la modelización de la formación del precio en mercados altamente fragmentados.

(d) Generalización de técnicas y modelos: Datos no estructurados de la misma

[24] Como notables excepciones, véase Kwan *et al.* (2024), que estudian variables condicionantes que afectan al proceso de formación del precio, y Akey *et al.* (2022) que estudia negociación informada.

naturaleza pero de fuentes diferentes pueden llevar a conclusiones también distintas. Como ejemplo tenemos los indicadores de sentimiento obtenidos a partir del análisis de noticias en prensa pero publicadas en periódicos con diferente orientación política, o medidas de divergencia de opinión calculadas a partir de Twitter versus Reddit. Además, los mercados financieros están en continua evolución y sujetos a cambios regulatorios. De hecho, no hay dos mercados financieros cuyas microestructuras sean exactamente iguales. Conclusiones derivadas de indicadores calculados para estudiar un determinado mercado podrían no ser generalizables a otros mercados. Por ejemplo, considere nuevamente el caso de índices de sentimiento calculados a partir de una determinada red social pero en este caso para diferentes países con diferentes niveles de cultura financiera.

(e) Mayor marco teórico: El éxito de muchas técnicas de ML, como las redes neuronales profundas, resulta de generar modelos con complejas relaciones e interacciones entre variables que en muchas ocasiones son difíciles de interpretar. Es por tanto fundamental desarrollar modelos teóricos que nos ayuden a entender las conclusiones derivadas de los modelos de ML, de manera que los mecanismos tras estas predicciones sean comprensibles para los académicos y expertos financieros.

(f) Colaboración interdisciplinar: El BD y los modelos de ML, especialmente los de alta complejidad, requieren una gran cantidad de recursos computacionales. Además, el investigador medio en finanzas no está probablemente entrenado o preparado para sacar ventaja de las oportunidades que las técnicas de ML ofrecen. Esta situación sugiere la necesidad de una mayor colaboración entre áreas de especialización para que la investigación en Mercados Financieros acabe de integrar plenamente estas técnicas y superar los retos que implican. Otra posibilidad es que el investigador en Mercados Financieros externalice el tratamiento de datos no estructurados, como en el caso de Chen *et al.* (2023), que usan *Large Language Models*, o Benamar *et al.* (2021) que miden la demanda de información y la incertidumbre utilizando datos de navegación proporcionados por un proveedor que transforma datos no estructurados en datos estructurados. El uso de proveedores especializados es acorde con nuestra visión de que un característica fundamental que los datos deben cumplir para favorecer el desarrollo científico es que éstos estén a disposición de otros investigadores (a un coste razonable) para facilitar la replicabilidad.

Las diferentes ramas de la ciencia avanzan y se enriquecen mutuamente. Hemos visto como los avances en computación, recogida de datos, y métodos algorítmicos que engloban la BD, IA, y ML ayudan y abren nuevas vías de conocimiento en el estudio de los Mercados Financieros. No obstante, estos avances suponen nuevos retos a la hora de adaptarnos a los cambios que éstos implican, integrarlos dentro del conocimiento ya establecido, y entender sus ventajas y desventajas. Hay que seguir subiendo la escalera del conocimiento peldaño a peldaño.

Referencias

Ackert, L. F., Jiang, L., Lee, H. S., y Liu, J. (2016). Influential investors in online stock forums. *International Review of Financial Analysis*, 45:39–46.

Akey, P., Grégoire, V., y Martineau, C. (2022). Price revelation from insider trading: Evidence from hacked earnings news. *Journal of Financial Economics*, 143(3):1162–1184.

Akramov, A., Kaniel, R., Kondor, P., y Sagi, J. (2021). Postfundamentals price drift in capital markets. *Management Science*, 68(10):7658–7681.

Andersen, T. G. (2000). Some reflections on analysis of high-frequency data. *Journal of Business & Economic Statistics*, 18(2):146–153.

Aquilina, M., Budish, E., y O'Neill, P. (2022). Quantifying the high-frequency trading "arms race": A simple new methodology and estimates. *The Quarterly Journal of Economics*, 137(1):493–564.

Aubry, M., Kräussl, R., Manso, G., y Spaenjers, C. (2023). Biased auctioneers. *The Journal of Finance*, 78(2):795–833.

Baker, M. y Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680.

Baker, M. y Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2):129–152.

Barber, B. M., Huang, X., Odean, T., y Schwarz, C. (2022). Attention-induced trading and returns: Evidence from robinhood users. *Journal of Finance*, 77(6):3141–3190.

Baron, M., Brogaard, J., Hagstr.ºmer, B., y Kirilenko, A. (2019). Risk and return in high-frequency trading. *Journal of Financial and Quantitative Analysis*, 54(3):993–1024.

Bartlett, R. P., Mccrary, J., y O'Hara, M. (2024). Tiny trades, big questions: Fractional shares. *Journal of Financial Economics*, 157.

Bekiros, S. D. (2013). Irrational fads, short-term memory emulation, and asset predictability. *Review of Financial Economics*, 22(4):213–219.

Benamar, H., Foucault, T., y Vega, C. (2021). Demand for information, uncertainty, and the response of us treasury securities to news. *The Review of Financial Studies*, 34(7):3403–3455.

Berkman, H., Dimitrov, V., Jain, P. C., Koch, P. D., y Tice, S. (2009). Sell on the news: Differences of opinion, short-sales constraints, and returns around earnings announcements. *Journal of Financial Economics*, 92(3):376–399.

Biais, B., Hillion, P., y Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the paris bourse. *The Journal of Finance*, 50(5):1655–1689.

Bloomberg News (2019). Jpmorgan commits hedge fund to ai in technology arms race.

Boehmer, E., Jones, C. M., Zhang, X., y Zhang, X. (2021). Tracking retail investor activity. *The Journal of Finance*, 76(5):2249–2305.

Brogaard, J., Hendershott, T., y Riordan, R. (2014). High-frequency trading and price discovery. *The Review of Financial Studies*, 27(8):2267–2306.

Brogaard, J., Hendershott, T., y Riordan, R. (2019). Price discovery without trading: Evidence from limit orders. *The Journal of Finance*, 74(4):1621–1658.

Brogaard, J. y Zareei, A. (2023). Machine learning and the stock market. *Journal of Financial and Quantitative Analysis*, 58(4):1431–1472.

Brown, S., Goetzmann, W., y Kumar, A. (1998). The dow theory: William peter hamilton's track record reconsidered. *Journal of Finance*, 53(4):1311–1333. 58th Annual Meeting of the American-Finance-Association, CHICAGO, ILLINOIS, JAN 03-05, 1998.

Bryzgalova, S., Huang, J., y Julliard, C. (2023). Bayesian solutions for the factor zoo: We just ran two quadrillion models. *The Journal of Finance*, 78(1):487–557.

Budish, E., Cramton, P., y Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, 130(4):1547–1621.

Buehler, H., Gonon, L., Teichmann, J., y Wood, B. (2019). Deep hedging. *Quantitative Finance*, 19(8):1271–1291.

Cao, C., Hansch, O., y Wang, X. (2008). The information content of an open limit-order book. *Journal of Financial and Quantitative Analysis*, 43(5):1023–1052.

Carlin, B., Olafsson, A., y Pagel, M. (2023). Mobile apps and financial decision making. *Review of Finance*, 27(3):977–996.

Carrion, A. (2013). Very fast money: High-frequency trading on the nasdaq. *Journal of Financial Markets*, 16(4):680–711.

Cartea, A., Arribas, I. P., y Sanchez-Betancourt, L. (2022a). Double-execution strategies using path signatures. *SIAM Journal of Financial Mathematics*, 13(4):1379–1417.

Cartea, A., Chang, P., Mroczka, M., y Oomen, R. (2022b). Ai-driven liquidity provision in otc financial markets. *Quantitative Finance*, 22(12):2171–2204.

Cartea, A., Chang, P., Penalva, J., y Waldon, H. (2022c). Algorithms can learn to collude: A folk theorem from learning with bounded rationality. *Available at SSRN 4293831*.

Cartea, A., Jaimungal, S., y Penalva, J. (2015). *Algorithmic and High-Frequency Trading*. Cambridge University Press.

Cartea, A., Payne, R., Penalva, J., y Tapia, M. (2019). Ultra-fast activity and intraday market quality. *Journal of Banking & Finance*, 99:157–181.

Chakrabarty, B., Comerton-Forde, C., y Pascual, R. (2024a). Identifying high frequency trading activity. *SSRN Working Paper*.

Chakrabarty, B., Hendershott, T., Nawn, S., y Pascual, R. (2024b). Order exposure in high frequency markets. *SSRN Working Paper*.

Chakrabarty, B., Pascual, R., y Shkilko, A. (2015). Informed trading and price discovery before corporate events. *Journal of Financial Markets*, 25:28–47.

Chatelais, N., Karamé, F., Patureau, L., y Topuz, S. (2023). Forecasting real activity. *Journal of International Money and Finance*, 123:102595.

Chen, H., De, P., Hu, Y., y Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5):1367–1403.

Chen, H., De, P., Hu, Y. J., y Hwang, B.-H. (2022). Investor attention and stock returns. *Journal of Financial and Quantitative Analysis*, 57(1):1–34.

Chen, Z., Kelly, B., y Xiu, D. (2023). Expected returns and large language models. *Working Paper*.

Chi, W., Huang, Y., Ke, B., y Verdi, R. (2023). The use of big data in finance: The case of financial analysts. *Working Paper*.

Chordia, T., Green, T. C., y Kottimukkalur, B. (2018). Macro news and micro news: Complements or substitutes? *The Review of Financial Studies*, 31(5):1723–1772.

Cochrane, J. H. (2009). *Asset Pricing: Revised Edition*. Princeton University Press.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, 66(4):1047–1108.

Collin-Dufresne, P. y Fos, V. (2015). Do prices reveal the presence of informed trading? *The Journal of Finance*, 70(4):1555–1582.

Cont, R., Kukanov, A., y Stoikov, S. (2014). The price impact of order book events. *Journal of Financial Econometrics*, 12(1):47–88.

Cookson, J. A., Engelberg, J., y Mullins, W. (2020). Does partisanship shape investor beliefs? evidence from the covid-19 pandemic. *The Review of Asset Pricing Studies*, 10(4):863–893.

Cookson, J. A., Engelberg, J., y Mullins, W. (2022). Echo chambers. *The Review of Financial Studies*, 35(5):2131–2178.

Cookson, J. A. y Niessen-Ruenzi, A. (2020). Why don't we agree? evidence from a social network of investors. *The Journal of Finance*, 75(1):173–228.

Da, Z., Engelberg, J., y Gao, P. (2015). The sum of all fears investor sentiment and asset prices. *The Review of Financial Studies*, 28(1):1–32.

Davis, J., Aliaga-D'ıaz, R., Shanahan, K., y Corr, A. (2020). Firm-level risk exposures and stock returns in the wake of covid-19. *Working Paper*.

De Spiegeleer, J., Madan, D. B., Reyners, S., y Schoutens, W. (2018). Machine learning for quantitative finance: fast derivative pricing, hedging and fitting. *Quantitative Finance*, 18(10):1635–1643.

Dello Preite, M., Uppal, R., Zaffaroni, P., y Zviadadze, I. (2024). Cross-sectional asset pricing with unsystematic risk. *SSRN*.

DeMiguel, V., Gil-Bazo, J., Nogales, F. J., y Santos, A. A. (2023). Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics*, 150(3):103737.

Deng, X., Guo, R., Lev, B., y Zhou, N. (2023). Seeing is believing: Annual report graphicity and stock returns. *Working Paper*.

Dessaint, O., Foucault, T., Frésard, L., y Matray, A. (2024). Does alternative data improve financial forecasting? the horizon effect. *The Journal of Finance*, 32(7):2625–2672.

Di Maggio, M., Kermani, A., y Majlesi, K. (2020). Stock market returns and consumption. *The Journal of Finance*, 75(6):3175–3219.

Dugast, J. y Foucault, T. (2018). Data abundance and asset price informativeness. *Journal of Financial Economics*, 130(2):367–391.

Easley, D., Kiefer, N. M., y O'Hara, M. (1997). One day in the life of a very common stock. *The Review of Financial Studies*, 10(3):805–835.

Easley, D., Kiefer, N. M., O'Hara, M., y Paperman, J. B. (1996). Liquidity, information, and infrequently traded stocks. *The Journal of Finance*, 51(4):1405–1436.

Easley, D., López de Prado, M., O'Hara, M., y Zhang, Z. (2021a). Microstructure in the machine age. *The Review of Financial Studies*, 34(7):3316–3363.

Easley, D., López de Prado, M., O'Hara, M., y Zhang, Z. (2021b). Microstructure in the machine age. *The Review of Financial Studies*, 34(7):3316–3363.

Eaton, G. W., Green, T. C., Roseman, B. S., y Wu, Y. (2022). Retail trader sophistication and stock market quality: Evidence from brokerage outages. *Journal of Financial Economics*, 146(2):502–528.

Edmans, A., Garel, A., Indriawan, I., y Krajcova, J. (2022). Music sentiment and stock returns around the world. *Journal of Financial Economics*, 145(2):234–254.

Engle, R. (2000). The econometrics of ultra-high-frequency data. *Econometrica*, 68(1):1–22.

Engle, R. F. y Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, pp. 1127–1162.

Fedyk, A. (2023). Front-page news: The effect of news positioning on financial markets. *The Journal of Finance*, 79(1):5–33.

Foucault, T., Pagano, M., y Roëll, A. (2024). *Market Liquidity: Theory, Evidence, and Policy*. Oxford University Press.

Froot, K., Kang, N., Ozik, G., y Sadka, R. (2017). What do measures of real-time corporate sales say about earnings surprises and post-announcement returns? *Journal of Financial Economics*, 125(1):143–162.

García, D., Herskovic, B., y Palacios, L. (2023). The colour of finance words. *Journal of Financial Economics*, 147(3):525–549.

Gargano, A. y Rossi, A. G. (2024). Goal setting and saving in the fintech era. *Journal of Finance*, 79(3):1931–1976.

Gentzkow, M., Kelly, B., y Taddy, M. (2019). Survey text as data. *Journal of Economic Literature*, 57(3):535–74.

Gerker, G. y Painter, P. (2022). The value of differing points of view. *The Review of Financial Studies*, 36(2):409–449.

Giannini, R., Irvine, P., y Shu, T. (2019). Convergence and divergence of investors' opinions. *Journal of Financial Markets*, 44:61–79.

Gibbons, B., Iliev, P., y Kalodimos, J. (2021). Analyst information acquisition via edgar. *Management Science*, 67(2):769–793.

Giglio, S., Liao, G., y Xiu, D. (2021). Risk and return in high-frequency trading. *The Review of Financial Studies*, 34(5):2223–2273.

Glaeser, E. L., Kincaid, M. K., y Naik, N. (2018). Computer vision and real estate: Do looks matter and do incentives determine looks? *Working Paper*.

Goldstein, I., Spatt, C. S., y Ye, M. (2021). Big data in finance. *The Review of Financial Studies*, 34(7):3213–3225.

Goodhart, C. A. y O'Hara, M. (1997). High frequency data in financial markets: Issues and applications. *Journal of Empirical Finance*, 4(2-3):73–114.

Griffin, J. M., Shu, T., y Topaloglu, S. (2023). Do municipal bond dealers give their customers "fair and reasonable"pricing? *Journal of Finance*, 78(2):887–934.

Gu, S., Kelly, B., y Xiu, D. (2020a). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Gu, S., Kelly, B., y Xiu, D. (2020b). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5, SI):2223–2273.

Harvey, C. R., Liu, Y., y Zhu, H. (2016). The robustness of prominent stock returns predictors. *The Review of Financial Studies*, 29(1):69–103.

Hasbrouck, J. (2007). *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press.

Hasbrouck, J. (2018). High frequency quoting: Short-term volatility in bids and offers. *Journal of Financial and Quantitative Analysis*, 53(2):613–641.

Hasbrouck, J. (2021). Price discovery in high resolution. *Journal of Financial Econometrics*, 19(3):395–430.

Hasbrouck, J. y Saar, G. (2013). Low-latency trading. *Journal of Financial Markets*, 16(4):646–679.

Hillert, A., Jacobs, H., y Müller, S. (2014). Media makes momentum. *The Review of Financial Studies*, 27(12):3467–3501.

Ho, M., Sun, Z., y Xin, J. (2015). Weighted elastic net penalized mean-variance portfolio design and computation. *SIAM Journal on Financial Mathematics*, 6(1):1220–1244.

Hong Kong Institute for Monetary and Financial Research (HKIMR) (2021). Algorithmic and high-frequency trading in hong kong's equity market: Adoption, market impact and risk management. Technical report, Hong Kong Academy of Finance (AoF).

Horvath, B., Muguruza, A., y Tomas, M. (2021). Deep learning volatility: a deep neural network perspective on pricing and calibration in (rough) volatility models. *Quantitative Finance*, 21(1):11–27.

Huang, D., Jiang, F., Li, K., Tong, G., y Zhou, G. (2022). Scaled pca: A new approach to dimension reduction. *Management Science*, 68(3):1678–1695.

Huang, D., Jiang, F., Tu, J., y Zhou, G. (2015). Investor sentiment aligned. *The Review of Financial Studies*, 28(3):791–837.

Huang, R. D. y Stoll, H. R. (1997). Did nyse floor traders monitor the limit order book? *Journal of Financial and Quantitative Analysis*, 32(2):249–268.

Huang, S., Kelly, B., y Xiu, D. (2023). Angel investors. *Journal of Financial Economics*, 149(2):161–178.

Hussain, S. M., Ahmad, W., y Ahmed, R. Z. (2023). Word power: A new approach for content analysis. *Working Paper*.

Jegadeesh, N. y Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729.

Jiang, J., Kelly, B., y Xiu, D. (2023). (re-) imag (in) ing price trends. *The Journal of Finance*, 78(6):3193–3249.

Jurado, K., Ludvigson, S. C., y Ng, S. (2015). Measuring uncertainty. *The American Economic Review*, 105(3):1177–1216.

Kaniel, R., Krüger, P., Starks, L. T., y Tham, W. (2023). Machine learning the skills of mutual fund managers. *Journal of Empirical Finance*, 68:1–22.

Kelly, B., Malamud, S., y Pedersen, L. H. (2023). Principal portfolios. *The Journal of Finance*, 78(1):347–387.

Kelly, B., Malamud, S., y Zhou, K. (2024). The virtue of complexity in return prediction. *The Journal of Finance*, 79(1):459–503.

Kelly, B. y Pruitt, S. (2013). Market expectations in the cross-section of present values. *The Journal of Finance*, 68(5):1721–1756.

Kelly, B. y Xiu, D. (2023). Financial machine learning. *Foundations and Trends in Finance*, 13(3-4):205–363.

Kercheval, A. N. y Zhang, Y. (2015). Modelling high-frequency limit order book dynamics with support vector machines. *Quantitative Finance*, 15(8, SI):1315–1329.

Kim, S. y Kim, S. (2020). Index tracking through deep latent representation learning. *Quantitative Finance*, 20(4):639–652.

Kolm, P. N., Turiel, J., y Westray, N. (2023). Deep order flow imbalance: Extracting alpha at multiple horizons from the limit order book. *Mathematical Finance*, 33(4):1044–1081.

Kwan, A., Philip, R., y Shkilko, A. (2024). The conduits of price discovery: A machine learning approach. *Working Paper*.

Ledoit, O. y Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. y Wolf, M. (2017). Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks. *The Review of Financial Studies*, 30(12):4349–4388.

Ledoit, O. y Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics*, 48(5):3043–3065.

Li, K., Mai, F., Shen, R., y Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34(7):3265–3315.

Light, N., Maslov, D., y Rytchkov, O. (2017). Aggregation of information about the cross section of stock returns: A latent variable approach. *Review of Financial Studies*, 30(4):1339–1381.

Lin, X. H., Cong, W., Feng, G., y He, J. (2023). Asset pricing with panel trees under global split criteria. *Working Paper*.

Loughran, T. y McDonald, B. (2020). Textual analysis in finance: A survey. *Annual Review of Financial Economics*, 12:299–319.

Ludvigson, S. C. y Ng, S. (2007). The empirical risk–return relation: a factor analysis approach. *Journal of Financial Economics*, 83(1):171–222.

Ludvigson, S. C. y Ng, S. (2010). A factor analysis of bond risk premia. En *Handbook of empirical economics and finance*, pp. 313–372. CRC Press.

Luss, R. y D'Aspremont, A. (2015). Predicting abnormal returns from news using text classification. *Quantitative Finance*, 15(6, SI):999–1012.

Makinen, Y., Kanniainen, J., Gabbouj, M., e Iosifidis, A. (2019). Forecasting jump arrivals in stock prices: new attention-based network architecture using limit order book data. *Quantitative Finance*, 19(12):2033–2050.

Manela, A. y Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1):137–162.

Marr, B. (2019). The revolutionary way of using artificial intelligence in hedge funds: The case of aidyia. *Forbes*.

Martin, I. y Nagel, S. (2022). Asset pricing with heterogeneous beliefs. *Journal of Financial Economics*, 144(3):605–636.

Mayew, W. J. y Venkatachalam, M. (2012). The power of voice: Managerial affective states and future firm performance. *The Journal of Finance*, 67(1):1–43.

Mittnik, S., Robinzonov, N., y Spindler, M. (2015). Stock market volatility - identifying major drivers. *Journal of Banking & Finance*, 58:1–14.

Muggleton, S. (2014). Machine learning, social learning and the discovery of the pulsars. *Nature Reviews Physics*, 1(3):201–209.

Murphy, K. P. (2022). *Probabilistic machine learning: an introduction*. MIT press.

Næs, R. y Skjeltorp, J. A. (2006). Order book characteristics and the volume–volatility relation: Empirical evidence from a limit order market. *Journal of Financial Markets*, 9(4):408–432.

Obaid, B., Puckett, A., y Yan, X. S. (2022). A picture is worth a thousand words: Textual analysis and mutual fund investment decisions. *Journal of Financial Economics*, 145(2):618–641.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, 116(2):257–270.

Pascual, R. y Veredas, D. (2010). Does the open limit order book matter in explaining informational volatility? *Journal of Financial Econometrics*, 8(1):57–87.

Penalva, J. S. y Tapia, M. (2021). Heterogeneity and competition in fragmented markets: Fees vs speed. *Applied Mathematical Finance*, 28(2):143–177.

Sautner, Z., Van Lent, L., Vilkov, G., y Zhang, R. (2023). Firm-level climate change exposure. *The Journal of Finance*, 78(3):1449–1498.

Shkilko, A. y Sokolov, K. (2020). Every cloud has a silver lining: Fast trading, microwave connectivity, and trading costs. *The Journal of Finance*, 75(6):2899–2927.

Sirignano, J. A. (2019). Deep learning for limit order books. *Quantitative Finance*, 19(4):549–570.

Stoll, H. R. (2002). Presidential address: friction. *The Journal of Finance*, 57(4):1479–1514.

Tashiro, D., Matsushima, H., Izumi, K., y Sakaji, H. (2019). Encoding of high-frequency order information and prediction of short-term stock price by deep learning. *Quantitative Finance*, 19(9, SI):1499–1506.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Tobeck, T. y Hronec, M. (2021). Does it pay to follow anomalies-research. *Journal of Financial Markets*, 56:100588.

U.S. Securities and Exchange Commission (2020). Policy Challenges and Research Opportunities in the Era of Big Data. https://www.sec.gov/newsroom/speeches-statements/policy-challenges-research-opportunities-era-big-data.

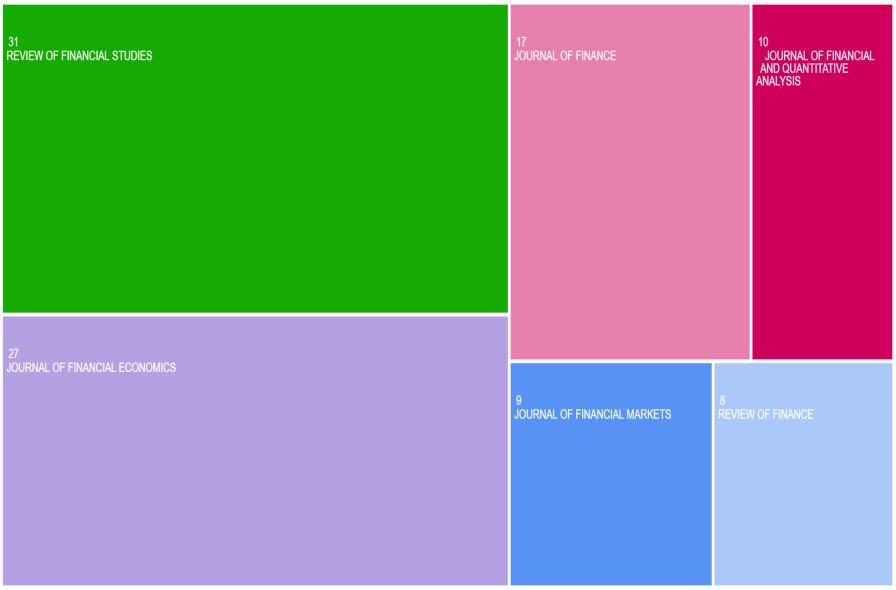
Wang, H. y Zhou, X. Y. (2020). Continuous-time mean-variance portfolio selection: A reinforcement learning framework. *Mathematical Finance*, 30(4):1273–1308.

Welch, I. (2022). The wisdom of the robinhood crowd. *Journal of Finance*, 77(3):1489–1527.

Wiese, M., Knobloch, R., Korn, R., y Kretschmer, P. (2020). Quant gans: deep generation of financial time series. *Quantitative Finance*, 20(9):1419–1440.

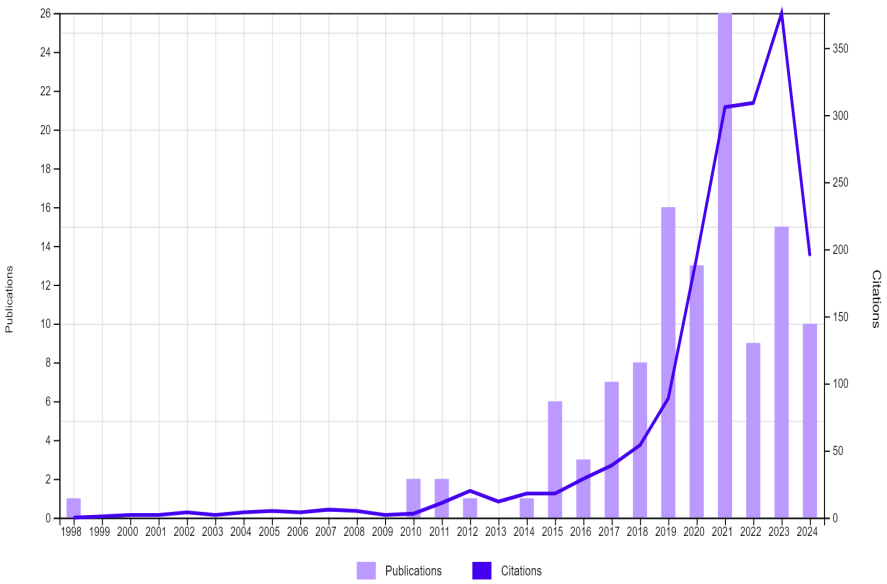
Zuckerman, G. y Hope, B. (2017). The quants run wall street now. *The Wall Street Journal*, (May 21).

Figura 1: IA en revistas de Finanzas generalistas (publicaciones)



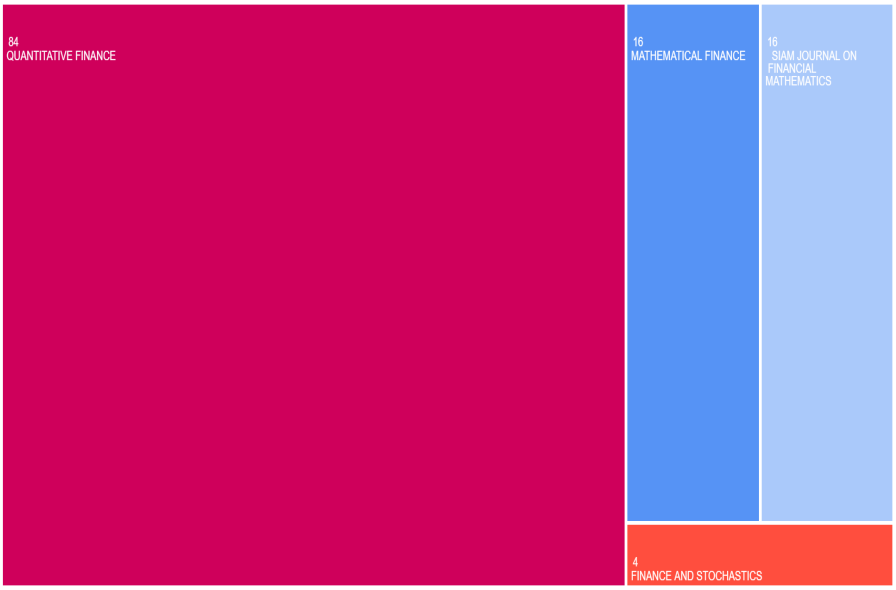
Fuente: Web Of Science (WoS)

Figura 2: IA en revistas de Finanzas generalistas (citas)



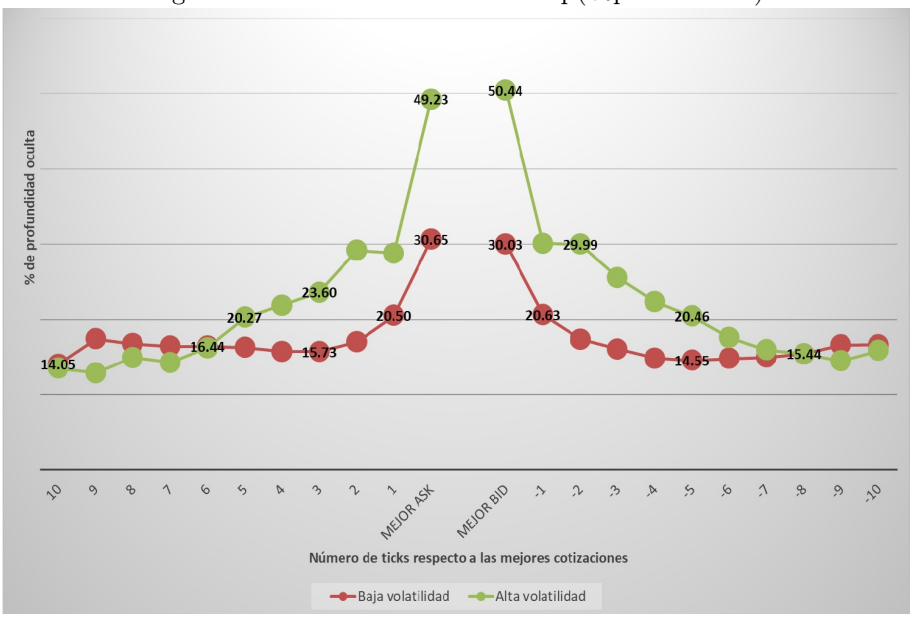
Fuente: Web Of Science (WoS)

Figura 3: IA en revistas de Finanzas Cuantitativas (publicaciones)



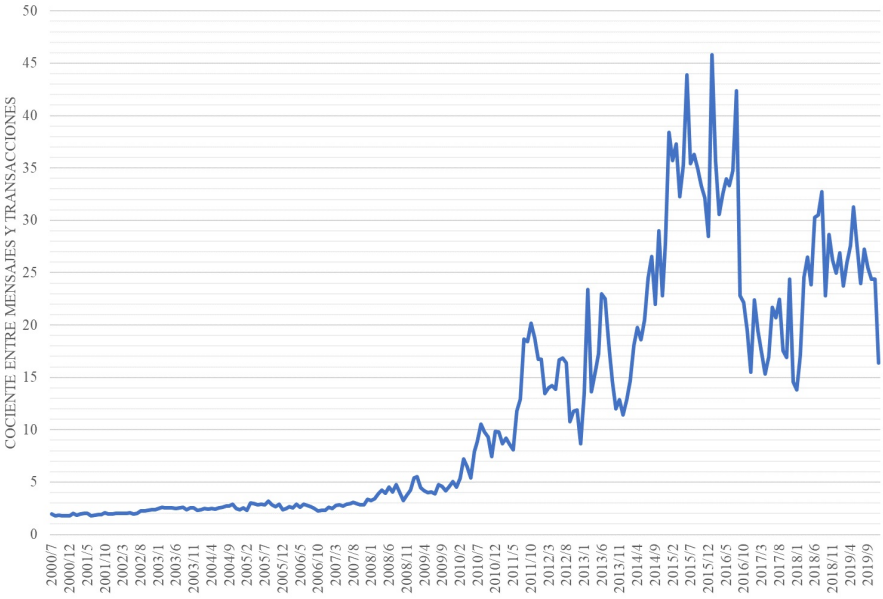
Fuente: Web Of Science (WoS)

Figura 4: El libro oculto del Nasdaq (%profundidad)



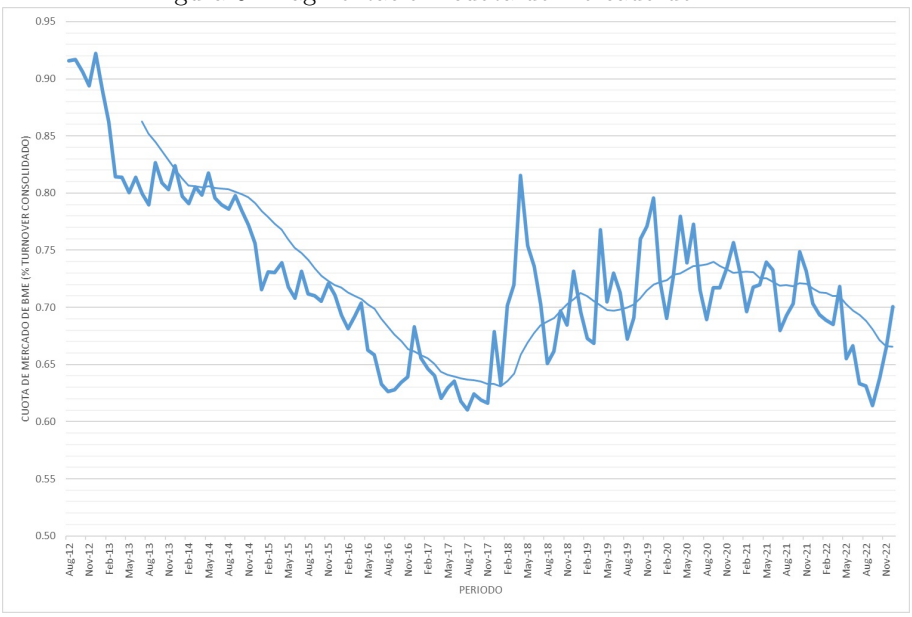
Fuente: Elaboración propia

Figura 5: Tráfico de Mensajes para TEF (2001-2019)



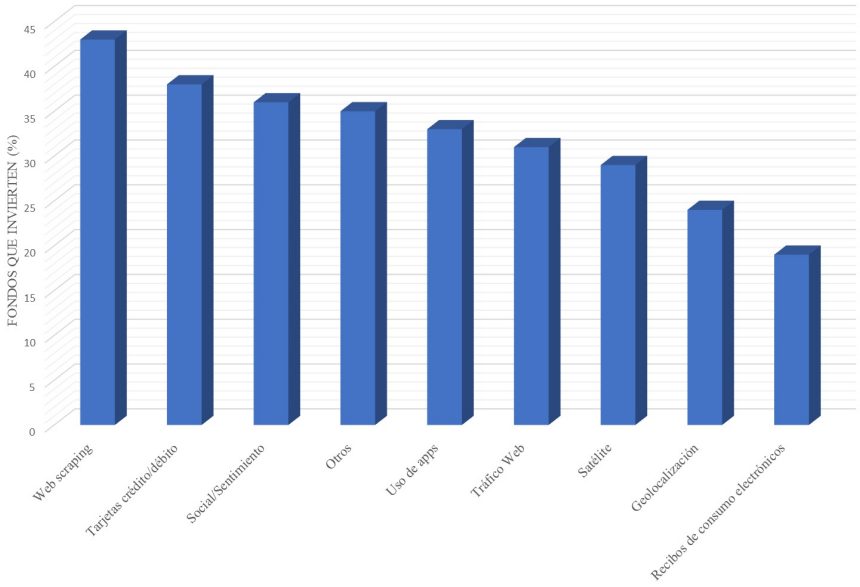
Fuente: Elaboración propia

Figura 6: Fragmentación: cuota de mercado de BME



Fuente: Elaboración propia

Figura 7: Datos alternativos



Fuente: www.alternativedata.org

Tabla I: Datos alternativos

Tipología de datos alternativos más comúnmente utilizados por la industria financiera. Fuente: https://alternativedata.org/.

Grupo	**Tipo**	**Descripción**
Individuos	Sentimiento	Datos obtenidos del procesamiento de texto en redes sociales, noticias, comunicaciones de la gerencia, y otras fuentes (p. ej., medición de sentimiento, viralidad de la marca, éxito publicitario, etc.)
	Tráfico Web	Datos sobre la cantidad, demografía, e historial de los usuarios que visitan un determinado sitio web.
	Aplicaciones	Datos sobre uso y evaluaciones de aplicaciones informáticas (p. ej., páginas de apuestas, de entrega de comida, servicios de *streaming* etc.)
	Encuestas	Datos recopilados de encuestas. Medición directa de sentimiento (p. ej., preferencia de marca; comportamiento del consumidor)
Negocios	Tarjetas	Datos de transacciones generados a partir de tarjetas de crédito y débito, con paneles estables de participantes.
	Web *scrapping*	Datos extraídos de sitios web públicos.
	Consumo	Datos de transacciones generados a partir de recibos electrónicos o *e-receipts*, i.e., documentos digitales que sirven como prueba de una transacción o compra.
	Otros	Otros conjuntos de datos alternativos, como los de punto de venta, los de gasto en publicidad etc.
Sensores	Geolocalización	Datos de tráfico peatonal disponibles a partir de señales WiFi o balizas Bluetooth (p. ej., alrededor de comercios, oficinas de bancos, sedes de empresas etc.)
	Satélite	Datos recopilados de satélites o drones de baja altitud; procesamiento de imágenes (p. ej., datos de estacionamiento; cadena de suministro; explotaciones agrícolas; construcción; producción/almacenamiento de petróleo y gas).
	Clima	Datos sobre patrones climáticos recopilados mediante sensores.
	Datos públicos	Datos de recursos públicos recopilados, agregados y manejables (p. ej., informes depositados en la SEC, datos de patentes, contratos gubernamentales, datos de importación/exportación, etc.)

Tabla II: Datos alternativos: Análisis de texto

Selección de trabajos académicos sobre mercados financieros que utilizan análisis lingüístico. En el Panel A, listamos trabajos que utilizan textos de redes sociales. En el panel B, escogemos trabajos que utilizan otros tipos de texto.

Tipo	**Referencia**	**Qué se pretende**
Panel A: Análisis de texto - Redes sociales		
(a) Mensajes cortos (StockTwits)	Dessaint *et al.* (2024)	Medir la disponibilidad de información orientada hacia el corto plazo.
Mensajes cortos (StockTwits)	Cookson *et al.* (2020)	Medir la orientación política del inversor.
(b) Mensajes cortos (StockTwits)	Cookson *et al.* (2022)	Identificar clusters de sentimiento.
(c) Mensajes cortos (StockTwits)	Cookson y Niessen-Ruenzi (2020)	Medir divergencias de opinión.
(d) Mensajes cortos (Seeking Alpha)	Chen *et al.* (2014)	Medir el contenido informativo en las opiniones vertidas en foros especializados.
(e) Mensajes cortos (Guba Eastmoney)	Ackert *et al.* (2016)	Medir el contenido informativo en las opiniones de inversores influyentes (con más seguidores)
Panel B: Análisis de texto - Otras fuentes		
(f) Informes de analistas (TR-Refinitiv)	Dessaint *et al.* (2024); Chi *et al.* (2023)	Evaluar el uso de datos alternativos por parte de analistas financieros.
(g) Informes anuales, noticias en prensa y transcripciones teleconferencias (WSJ, Seeking Alpha, Wall Street Horizons)	García *et al.* (2023)	Construir un diccionario de palabras con connotación positiva o negativa para medir sentimiento.
(h) Noticias (Meta, Google, OpenAI, TR-Refinitiv)	Chen *et al.* (2023)	Aplicar Modelos de Lenguaje de Gran Escala para medir sentimiento y predecir movimientos del mercado.
(i) Noticias de prensa (LexisNexis)	Hillert *et al.* (2014)	Construir una medida de desacuerdo entre periodistas.
(j) Noticias en terminales electrónicos (Bloomberg)	Fedyk (2023)	Medir el posicionamiento relativo de una noticia en el terminal.
(k) Noticias en prensa (WSJ)	Manela y Moreira (2017)	Construir medidas de incertidumbre basadas en el texto de noticias.
(l) Versión leíble por máquinas de informes anuales (EDGAR, SEC)	Davis *et al.* (2020)	Identificar factores de riesgo a nivel de empresa.
(m) Transcripciones informes anuales (Eikon)	Sautner *et al.* (2023)	Medir la atención mostrada por los participantes en *earnings conference calls* a la exposición de la empresa al riesgo climático.
(n) Anuncios de beneficios. Fuente: I/B/E/S y Ravenpack	Akey *et al.* (2022)	Medir negociación informada.

Tabla III: Datos alternativos: Análisis de imagen, sonido y otros

Selección de trabajos académicos sobre mercados financieros que utilizan datos no estructurados que no son texto. En el Panel A, listamos trabajos que utilizan análisis de imagen y sonido. En el panel B, escogemos trabajos que utilizan vídeo, actividad en línea y otros tipos de datos alternativos.

Tipo	**Referencia**	**Qué se pretende**
Panel A: Análisis de imagen y sonido		
(a) Imagenes satelitales de zonas de aparcamiento (Orbital Insights)	Gerker y Painter (2022)	Medición del performance local de empresas.
(b) Gráficos de precios	Jiang *et al.* (2023)	Extraer los patrones de precios con mayor poder predictivo.
(c) Gráficos en la versión .ªmigable"de informes anuales (páginas web de empresas)	Deng *et al.* (2023)	Identificar la presencia de gráficos en los informes anuales.
(d) Fotos en prensa (WSJ)	Obaid *et al.* (2022)	Construir un indicador de sentimiento ("foto pesimismo")
Panel B: Análisis de video, actividad en línea y otros		
(e) Canciones en repositorios de música (Spotify)	Edmans *et al.* (2022)	Construir un indicador de sentimiento a nivel nacional basado en la positividad de la música escuchada.
(f) Vídeos de presentaciones de startups (SharkTank; Startup Batterfield)	Huang *et al.* (2023)	Medir primeras impresiones en presentaciones de emprendedores.
(g) Uso de aplicaciones (EDGAR)	Gibbons *et al.* (2021)	Medir adquisición de información por analistas financieros.
(h) Búsquedas online (Google)	Da *et al.* (2015)	Medir sentimiento
(i) Registros de cajas registradoras (Statistics Sweden)	Di Maggio *et al.* (2020)	Estudiar las decisiones financieras de los hogares
(j) Dispositivos electrónicos (teléfonos móviles, tabletas) (MKT-Mediastats)	Froot *et al.* (2017)	Medir actividad de consumo por parte de las familias.

Tabla IV: Métodos de ML en la literatura de Mercados Financieros

Selección de trabajos académicos que utilizan los principales métodos de ML utilizados en la literatura en Mercados Financieros.

Método	**Referencia**	**Aplicación ilustrativa**
Panel A: *Representation learning*		
Principal Components	Ludvigson y Ng (2007)	Utiliza PCA para estimar la estructura factorial en la valoración de activos.
Partial Least Squares	Chen *et al.* (2022)	Utiliza PLS para extraer la información conjunta de diferentes medidas de la atención de los inversores.
SVM	Luss y D'Aspremont (2015)	Utiliza SVM para clasificar noticias y predecir grandes movimientos en precios de acciones.
Árboles de decisión	Lin *et al.* (2023)	Utiliza árboles de decisión en datos de panel para extraer el factor de descuento estocástico para valorar activos financieros.
Multiple Regression Methods	Tobeck y Hronec (2021)	Utiliza *Random Forests*, *GLR*, y *Neural Nets* para evaluar la rentabilidad de estrategias basadas en errores de valoración.
Panel B: Redes Neuronales		
Deep NN	Horvath *et al.* (2021)	Utiliza redes neuronales profundas para modelar la superficie de volatilidad implícita de opciones.
Convolutional NN	Jiang *et al.* (2023)	Utiliza CNN como método de análisis técnico de series de precios para predecirlos.
Decoder	Kolm *et al.* (2023)	Modelización de libro de órdenes para generar estrategias que anticipan precios.
Panel C: Regularización		
Shrinkage	Kelly *et al.* (2024)	Modelo teórico que utiliza la teoría de matrices aleatorias para justificar la superioridad estadística de métodos de *shrinkage* que muchas variables.
	DeMiguel *et al.* (2023)	Utiliza métodos de ML para estimar el rendimiento esperado de fondos de inversión.
Panel D: *Reinforcement Learning*		
RL	Cartea *et al.* (2022c)	Modelo teórico de cómo algoritmos de RL aprenden y compiten de manera conjunta.
	Buehler *et al.* (2019)	Optimización de carteras de derivados.