TRABAJO DE INVESTIGACIÓN FINAL

Humanos versus máquinas en la gestión de carteras: una comparación en el mercado argentino durante episodios de estrés financiero (2019-2026)

Autor/es:

Matteo, Julián LU: 1167139

Szulman, Sebastián Mauricio LU: 1131145

Carrera:

Lic. Finanzas

Tutor/a:

Favata, Federico

Año:

2026

Universidad Argentina de la Empresa
Facultad de Ciencias Económicas

UADE

Agradecimientos:

En primer lugar, queremos agradecer a nuestros padres y familias, quienes nos acompañaron y sostuvieron a lo largo de toda la carrera universitaria, brindándonos el apoyo, la paciencia y la confianza necesarios para poder dedicarnos a nuestra formación y llegar hasta esta instancia final.

Asimismo, queremos expresar nuestro agradecimiento a la profesora Karina Díaz, por su guía, sus correcciones y el tiempo dedicado al seguimiento de este Trabajo de Investigación Final. Sus observaciones contribuyeron de manera significativa a mejorar la calidad del trabajo y a ordenar el proceso de investigación.

Agradecemos especialmente a nuestro tutor, Federico Favata, por su disposición, su compromiso y sus aportes a lo largo de todo el desarrollo de esta investigación.

También queremos agradecer a nuestros compañeros de facultad, con quienes compartimos años de estudio, esfuerzo y aprendizaje. Su apoyo, compañía e intercambio constante hicieron más llevadero y enriquecedor este recorrido académico.

Por último, agradecemos a nuestros amigos y a todas aquellas personas que, de una u otra manera, nos acompañaron durante este proceso, brindándonos apoyo, motivación y momentos de distensión que fueron fundamentales para llegar hasta este logro.

Resumen

El presente Trabajo de Investigación Final analiza el desempeño comparativo entre estrategias de gestión de carteras humana y algorítmica en el mercado argentino durante el período 2019–2026, integrando teoría financiera, finanzas conductuales y modelos de machine learning. El estudio se desarrolla en un contexto de mercado emergente caracterizado por volatilidad y restricciones cambiarias. La investigación adopta un enfoque cuantitativo, comparativo y explicativo, utilizando datos diarios de acciones argentinas, fondos comunes de inversión y precios en dólares CCL. Las estrategias algorítmicas se construyen mediante modelos Random Forest, XGBoost y LSTM, cuyas predicciones alimentan una optimización de carteras basada en Markowitz. Los resultados muestran que las estrategias algorítmicas superaron consistentemente a los fondos comunes de inversión tanto en el período completo (632% de retorno acumulado en dólares frente al 117%) como en los cuatro episodios de estrés financiero analizados, con una ventaja explicada por una recuperación más rápida tras los shocks.

Palabras clave: gestión de carteras; machine learning ; finanzas conductuales; mercado argentino; fondos comunes de inversión.

Abstract

This Final Research Project analyzes the comparative performance between human and algorithmic portfolio management strategies in the Argentine market during the 2019–2026 period, integrating financial theory, behavioral finance and machine learning models. The study is conducted in an emerging market context characterized by volatility and foreign exchange restrictions. The research follows a quantitative, comparative and explanatory approach, using daily data from Argentine equities, mutual funds and prices expressed in CCL dollars. Algorithmic strategies are built using Random Forest, XGBoost and LSTM models, whose predictions feed a Markowitz-based portfolio optimization process. The results show that algorithmic strategies consistently outperformed mutual funds, both over the complete period (632% cumulative return in dollars versus 117%) and across the four financial stress episodes analyzed, with the advantage mainly explained by a faster recovery following shocks.

Keywords: portfolio management; machine learning; behavioral finance; Argentine market; mutual funds.

Contenido

Introducción

El presente Trabajo de Investigación Final tiene como objetivo analizar el desempeño comparativo entre estrategias de gestión de carteras humana y algorítmica en el mercado argentino durante el período 2019–2026. La investigación parte del debate actual sobre la incorporación de modelos de machine learning en los procesos de inversión y su posible capacidad para mejorar el desempeño frente a la gestión activa tradicional. En particular, se busca evaluar si las estrategias basadas en modelos algorítmicos presentan un rendimiento superior al de los fondos comunes de inversión de renta variable argentina, especialmente en períodos de estrés financiero y elevada volatilidad.

A partir del cuestionamiento a la racionalidad plena asumida por los modelos financieros clásicos, como la Teoría Moderna de Portafolios de Markowitz y la Hipótesis de Mercados Eficientes de Fama, el trabajo integra distintos enfoques teóricos para abordar la comparación entre ambos tipos de gestión. En primer lugar, se desarrollan los fundamentos de la selección de carteras, la eficiencia de mercado y la búsqueda de alpha. Luego, se analizan las limitaciones de la gestión humana desde las finanzas conductuales, con foco en los sesgos cognitivos y emocionales que pueden afectar la toma de decisiones. A su vez, se estudian los modelos algorítmicos de inversión, sus capacidades predictivas y sus propias limitaciones metodológicas, incluyendo sesgos vinculados a datos, sobreajuste y cambios de régimen.

La investigación se enfoca en el mercado argentino por tratarse de un contexto emergente caracterizado por menor profundidad, restricciones cambiarias, volatilidad elevada y fuerte sensibilidad frente a eventos políticos y macroeconómicos. Para ello, se adopta una metodología cuantitativa, comparativa y explicativa basada en datos diarios de acciones argentinas y fondos comunes de inversión de renta variable, expresados en dólares CCL. Las estrategias algorítmicas se construyen mediante modelos Random Forest, XGBoost y LSTM, cuyas predicciones se utilizan como insumo para una optimización de carteras bajo el criterio de Markowitz. Los resultados permitirán aportar evidencia empírica sobre la capacidad relativa de la gestión humana y algorítmica para generar desempeño en un mercado emergente como el argentino.

Planteamiento del problema

En la gestión de carteras de inversión, la comparación entre estrategias de gestión activa humana y estrategias basadas en modelos algorítmicos presenta dificultades empíricas significativas, particularmente en el contexto de mercados emergentes. Según Bekaert y Harvey (2002), estos mercados suelen caracterizarse por menor profundidad financiera, mayores restricciones operativas y niveles más elevados de volatilidad respecto de los mercados desarrollados, lo que introduce distorsiones en los resultados y limita la posibilidad de realizar comparaciones directas y consistentes entre distintos enfoques de gestión. Estas características adquieren especial relevancia dentro del mercado argentino, donde las condiciones de ineficiencia relativa y sensibilidad frente a shocks económicos y políticos afectan significativamente el comportamiento de los activos financieros. Los principales modelos algorítmicos utilizados en esta investigación serán desarrollados posteriormente en el marco teórico correspondiente.

En este sentido, no se encuentra claramente establecido cuál de estos enfoques de gestión, gestión activa humana y gestión algorítmica basada en modelos de machine learning resulta más eficiente en términos de rendimiento. Mientras que los modelos algorítmicos permiten procesar grandes volúmenes de información de manera sistemática y disciplinada, la gestión humana incorpora juicio contextual y capacidad de adaptación frente a eventos extraordinarios. No obstante, Barber y Odean (2001) sostienen que las decisiones financieras humanas pueden verse afectadas por sesgos conductuales y limitaciones cognitivas que alteran el proceso racional de toma de decisiones. En la misma línea, Kahneman y Tversky (1979) señalan que las decisiones bajo incertidumbre suelen encontrarse influenciadas por heurísticas y sesgos cognitivos, mientras que Shefrin (2000) argumenta que factores emocionales y psicológicos afectan significativamente el comportamiento financiero de los inversores. Las características y limitaciones asociadas tanto a los modelos algorítmicos como a la gestión humana y su juicio contextual serán abordadas posteriormente en el desarrollo del marco teórico.

Frente a esta problemática, surge la necesidad de desarrollar un enfoque que permita aproximar dicha comparación de manera más controlada. En este marco, Gu et al. (2020) sostienen que el uso de modelos de machine learning permite construir estrategias sistemáticas y replicables, facilitando el análisis empírico del desempeño relativo entre gestión algorítmica y gestión humana, particularmente en períodos de estrés financiero dentro del mercado argentino.

La creciente incorporación de modelos de machine learning en la gestión de carteras ha generado un debate relevante dentro de las finanzas modernas respecto a su capacidad para mejorar el desempeño frente a la gestión activa tradicional. Krauss et al. (2017) y Gu et al.(2020) muestran que determinados modelos algorítmicos pueden capturar relaciones no lineales y mejorar la predicción de retornos financieros respecto de enfoques tradicionales. No obstante, la mayor parte de esta evidencia empírica se concentra en mercados desarrollados, particularmente en el mercado estadounidense, dejando un vacío respecto de su aplicabilidad en mercados emergentes, caracterizados por mayores niveles de ineficiencia, volatilidad y restricciones operativas (Bekaert & Harvey, 2002).

Siguiendo a Bekaert y Harvey (2002), analizar el comportamiento comparativo entre gestión algorítmica y gestión humana en períodos de estrés financiero resulta especialmente relevante, dado que en estos contextos los supuestos de eficiencia tienden a debilitarse y los sesgos conductuales pueden impactar con mayor intensidad sobre la toma de decisiones. Kahneman (2011) sostiene que los contextos de incertidumbre favorecen decisiones más intuitivas y menos racionales, amplificando el efecto de los sesgos cognitivos sobre el comportamiento financiero. Dentro de esta investigación, los principales períodos de estrés considerados corresponden a las PASO de 2019, la crisis derivada del COVID-19 en 2020, el rally financiero argentino de 2023 y el escenario electoral de 2025.

Si bien se ha avanzado en el análisis de estrategias cuantitativas, la mayor parte de estos estudios como el de Harvey et al. (2017) o Grobys et al.(2022) se concentran en mercados desarrollados, dejando un vacío relevante en contextos como el local. Actualmente, no existe evidencia empírica concluyente que permita comparar de manera consistente ambos enfoques en mercados emergentes como el argentino. Esta carencia no solo se debe a la falta de estudios, sino también a las dificultades metodológicas que implica aislar el efecto del tipo de gestión en entornos donde las condiciones de mercado introducen distorsiones.

Preguntas de investigación

¿Qué diferencias se observan en el desempeño entre carteras gestionadas mediante modelos de machine learning y aquellas administradas por gestores humanos en el

mercado argentino durante el período 2019–2026, particularmente en episodios de estrés financiero?

¿Qué sesgos conductuales pueden afectar la toma de decisiones de gestores activos y cómo se diferencian de las limitaciones y sesgos propios de los modelos de gestión algorítmica?

¿Qué rol cumplen los modelos de machine learning en la construcción de estrategias de inversión y en la generación de evidencia empírica que permita analizar y comparar el desempeño de la gestión algorítmica frente a la gestión humana en el mercado argentino?

¿Cómo influyen las condiciones del mercado argentino en la capacidad de comparar de manera consistente estrategias de gestión humana y algorítmica?

Justificación de la investigación

Desde el punto de vista teórico, la investigación contribuye a integrar distintos enfoques del campo financiero, finanzas conductuales, teoría de mercados eficientes y modelos de machine learning, aportando evidencia empírica en un contexto poco explorado como el mercado argentino. Esto permite ampliar el conocimiento sobre el desempeño de distintas estrategias de inversión en escenarios de alta incertidumbre, así como evaluar la capacidad de los modelos algorítmicos para capturar oportunidades de rendimiento frente a las limitaciones propias de la gestión humana.

En términos prácticos, los principales beneficiarios de esta investigación son los gestores de carteras, las instituciones financieras y los inversores institucionales, quienes podrán utilizar los resultados para evaluar la incorporación de herramientas algorítmicas en sus procesos de inversión y mejorar la asignación de capital en contextos adversos. Asimismo, el estudio aporta insumos para el desarrollo de estrategias híbridas que combinen análisis cuantitativo con juicio humano, contribuyendo a optimizar la toma de decisiones y a comprender en qué condiciones cada enfoque de gestión puede resultar más eficiente.

Objetivos

1. Analizar el desempeño entre estrategias de gestión de carteras humana y algorítmica en el mercado argentino durante el período 2019–

2026, evaluando el aporte de los modelos de machine learning como herramienta para su comparación en un contexto de mercado emergente.

2. Analizar los sesgos conductuales que afectan la toma de decisiones en la gesti ón activa, así como las limitaciones propias de los modelos algorítmicos, evalua ndo su impacto en el desempeño de las estrategias de inversión.
3. Desarrollar y aplicar modelos de machine learning para la construcción de estrategias de inversión, analizando su desempeño y compar ando su capacidad para generar evidencia empírica frente a la gestión activa humana.
4. Analizar la influencia de las condiciones estructurales del mercado argentino sobre la capacidad de comparar de manera consistente el desempeño entre estrategias de gestión humana y algorítmica.

Hipótesis

Se observan diferencias significativas a favor de las estrategias basadas en modelos de machine learning respecto de la gestión activa humana en el mercado argentino durante el período 2019–2026, particularmente en episodios de estrés financiero.

Marco teórico

1. Fundamentos teóricos: optimización de carteras y eficiencia de mercados

1.1 Selección de carteras y la frontera eficiente

La teoría moderna de carteras se origina en el trabajo seminal de Harry Markowitz, quien en 1952 publica en The Journal of Finance el artículo Portfolio Selection. Markowitz cuestiona la idea de maximizar el retorno esperado de una cartera, demostrando que esa regla conduciría a una solución concentrada en un único activo, y propone en su lugar la regla expected return-variance (E-V): el inversor evalúa cada cartera considerando simultáneamente el retorno esperado, deseable, y la varianza, indeseable. El autor sostiene que *"el inversor considera (o debería considerar) que la rentabilidad esperada es algo deseable y que la variabilidad de la rentabilidad es algo indeseable"* (Markowitz, 1952, p. 77). Markowitz (1952) demuestra que la varianza de una cartera

depende no solo de las varianzas individuales sino también de las covarianzas entre activos, de modo que la diversificación reduce el riesgo total: *"La diversificación es tanto una realidad observable como una medida sensata; cualquier norma de conducta que no implique la superioridad de la diversificación debe rechazarse tanto como hipótesis como máxima"* (Markowitz, 1952, p. 77). A partir de este planteo introduce la noción de cartera eficiente, aquella que minimiza la varianza para un retorno dado cuyo conjunto describe la frontera eficiente en el plano retorno-riesgo.

La relevancia de Markowitz (1952) para esta investigación es directa: los modelos de machine learning (Random Forest, XGBoost y LSTM) generan las predicciones de retornos esperados que alimentan día a día el problema de optimización, actuando como motores de pronóstico mientras la asignación de pesos se resuelve con la lógica media-varianza. Las estrategias algorítmicas no abandonan el marco teórico tradicional sino que lo extienden, reemplazando estimaciones discrecionales por predicciones de aprendizaje automático y conservando la frontera eficiente como criterio de asignación. Si los modelos producen estimaciones más precisas que las decisiones humanas que orientan a los FCI, las carteras resultantes deberían ubicarse sobre porciones más favorables de la frontera eficiente: sobre esta posibilidad se sostiene conceptualmente la hipótesis central del trabajo.

1.2 Hipótesis de mercados eficientes

Si bien Markowitz (1952) proporcionó el aparato matemático para construir carteras óptimas, su modelo no especifica cómo se forman los precios ni en qué medida la información disponible se incorpora a ellos. Esa pregunta es el núcleo de la Hipótesis de Mercados Eficientes (HME), formulada por Eugene Fama en su artículo Efficient Capital Markets: A Review of Theory and Empirical Work, publicado en The Journal of Finance en 1970. El autor sostiene que "*Un mercado en el que los precios 'reflejan plenamente' en todo momento la información disponible se denomina 'eficiente'."* (Fama, 1970, p. 383), lo que implica que ningún inversor puede obtener retornos superiores de manera sistemática utilizando información públicamente disponible y que la gestión activa no debería poder generar alpha de forma persistente.

Fama propone una clasificación en tres formas de eficiencia: la débil, donde los precios reflejan toda la información histórica de precios y volúmenes; la semi-fuerte, donde incorporan además toda la información pública disponible, incluyendo estados contables e indicadores macroeconómicos; y la fuerte, donde reflejan incluso la información

privilegiada, aunque esta última es generalmente rechazada por la evidencia. Fama concluye que los mercados desarrollados muestran buen ajuste a las formas débil y semi-fuerte, reconociendo que *"Las pruebas empíricas que respaldan el modelo de mercados eficientes son abundantes y (algo bastante singular en economía) las pruebas contradictorias son escasas"* (Fama, 1970, p. 416).

La hipótesis de superioridad del machine learning planteada en este trabajo supone, implícitamente, que el mercado argentino no se ajusta plenamente a la eficiencia semi-fuerte y que existen regularidades explotables: patrones temporales en los retornos, demoras en la incorporación de información o relaciones no lineales que exceden la capacidad del analista humano promedio. El presente trabajo no se propone refutar la HME sino evaluar si, bajo las condiciones del mercado argentino durante 2019–2026 controles cambiarios, menor profundidad y episodios recurrentes de estrés financiero, existen ineficiencias suficientes para que un sistema de machine learning pueda explotarlas mejor que los gestores humanos de los FCI locales. La HME funciona así como marco de contraste: cuanto más se aleje el desempeño algorítmico del esperado bajo eficiencia, mayor será la evidencia de ineficiencias explotables.

1.3 La paradoja de Grossman-Stiglitz: ¿por qué la búsqueda de alpha es sostenible?

La Hipótesis de Mercados Eficientes formulada por Fama (1970) deja planteada una tensión interna formalizada por Sanford Grossman y Joseph Stiglitz en On the Impossibility of Informationally Efficient Markets, publicado en la American Economic Review en 1980. La paradoja parte de una observación simple: si los precios reflejaran perfectamente toda la información disponible, los inversores no tendrían incentivos para producirla, pero sin inversores que produzcan información no existiría mecanismo para incorporarla a los precios. El núcleo es que la información es costosa: *"Dado que la información tiene un coste, los precios no pueden reflejar perfectamente la información disponible, ya que, de ser así, quienes invierten recursos para obtenerla no recibirían ninguna compensación."* (Grossman & Stiglitz, 1980, p. 393). Los autores demuestran formalmente la imposibilidad de mercados informacionalmente eficientes, concluyendo que los mercados reales deben presentar siempre un grado de ineficiencia residual suficiente para compensar a los inversores informados: *"existe un conflicto fundamental entre la eficacia con la que los mercados difunden la información y los incentivos para obtenerla"* (Grossman & Stiglitz, 1980, p. 405).

En mercados emergentes como el argentino, el costo de adquirir y procesar información tiende a ser mayor por la menor disponibilidad de datos, la fragmentación informativa y la menor cobertura analítica, lo que sugiere, bajo el marco de Grossman y Stiglitz (1980), un grado de ineficiencia residual también mayor. Esto ofrece un terreno más fértil para que un sistema capaz de procesar grandes volúmenes de datos de forma sistemática y a bajo costo marginal como un modelo de machine learning pueda capturar parte de esas ineficiencias. La cuestión empírica que esta tesis pone a prueba es precisamente si los modelos de machine learning son capaces de capturar más eficazmente esa ineficiencia residual que los gestores humanos de los FCI argentinos, pregunta cuya respuesta depende también de las capacidades, sesgos sistemáticos y desempeño documentado del decisor humano, que se abordan en el próximo capítulo.

2. La gestión humana: capacidades, sesgos y evidencia empírica

2.1 Finanzas conductuales y racionalidad limitada en la toma de decisiones financieras

Kahneman y Tversky (1979) sostienen que, al evaluar resultados, los individuos comparan lo obtenido con un punto de referencia psicológico previo en lugar de juzgarlo en términos absolutos, lo que da lugar a comportamientos que con frecuencia se desvían de los postulados de la racionalidad clásica. La teoría financiera tradicional se desarrolló sobre el supuesto de racionalidad plena de los agentes, postulado defendido por autores como Fama (1970), quien sostuvo que los precios reflejan adecuadamente la información existente en el mercado. Sin embargo, episodios de burbujas especulativas, crisis financieras y comportamientos colectivos irracionales evidenciaron que las decisiones de inversión no siempre responden a criterios estrictamente racionales, dando origen a las finanzas conductuales (behavioral finance), corriente que incorpora aportes de la psicología cognitiva para analizar cómo los factores emocionales influyen sobre la toma de decisiones financieras. Uno de sus principales referentes, Hersh Shefrin, las definió así*: "Las finanzas conductuales estudian cómo la psicología afecta a las finanzas. La psicología constituye la base de los deseos, motivaciones y objetivos humanos, así como también de numerosos errores derivados de ilusiones perceptivas, exceso de confianza y emociones."* (Shefrin, 2000, p. 4).

Las finanzas conductuales representan un marco teórico fundamental para comprender las limitaciones inherentes a la gestión activa humana. Si bien los gestores profesionales

poseen capacidad de análisis contextual e interpretación macroeconómica, sus decisiones continúan expuestas a mecanismos psicológicos y emocionales que pueden deteriorar el desempeño de las estrategias de inversión, cuestión central en el presente trabajo dado que uno de sus principales objetivos consiste en comparar dichas limitaciones frente a estrategias sistemáticas construidas mediante modelos de machine learning.

2.1.1 Teoría Prospectiva (Prospect Theory)

Kahneman y Tversky (1979) sostienen que las pérdidas generan un impacto psicológico mayor que las ganancias equivalentes. La Teoría Prospectiva (Prospect Theory), desarrollada por Kahneman y Tversky en 1979, surge como una crítica a la Teoría de la Utilidad Esperada (Expected Utility Theory), modelo normativo que describe cómo deberían tomar decisiones los individuos racionales. En contraste, la Teoría Prospectiva se presenta como un modelo descriptivo orientado a explicar cómo las personas realmente toman decisiones bajo incertidumbre, representando uno de los principales quiebres conceptuales respecto de las finanzas tradicionales al introducir formalmente el análisis psicológico en el estudio de las decisiones económicas y financieras.

Kahneman y Tversky (1979) argumentan que los individuos no evalúan resultados finales absolutos sino cambios relativos respecto de un punto de referencia psicológico: las personas perciben ganancias y pérdidas comparando continuamente los resultados con expectativas previas, precios históricos o niveles de riqueza de referencia, punto que no es fijo sino que puede verse afectado por emociones, contexto económico y experiencias recientes.

La Teoría Prospectiva sentó las bases para el desarrollo posterior de numerosos estudios sobre sesgos cognitivos aplicados a inversiones, incluyendo el exceso de confianza (overconfidence bias), el efecto disposición (disposition effect), el sesgo de confirmación (confirmation bias), el sesgo de anclaje (anchoring bias), el comportamiento de manada (herding behavior) y la contabilidad mental (mental accounting). En consecuencia, representa uno de los pilares fundamentales de las finanzas conductuales modernas, al demostrar que las decisiones financieras no dependen exclusivamente de cálculos racionales de utilidad esperada sino también de factores psicológicos, emocionales y contextuales.

2.1.2 Procesamiento dual de información: Sistema 1 y Sistema 2

A partir de los desarrollos de las finanzas conductuales, Daniel Kahneman (2011) formuló la teoría del procesamiento dual de información para explicar cómo los individuos toman decisiones bajo condiciones de incertidumbre. Según el autor, el razonamiento humano se estructura sobre la interacción de dos sistemas cognitivos: el Sistema 1, que opera de manera rápida, automática e intuitiva, y el Sistema 2, que funciona de forma deliberativa, racional y analítica. Kahneman (2011) sostiene que gran parte de las decisiones cotidianas son dominadas por el Sistema 1, ya que los individuos tienden naturalmente a minimizar el esfuerzo mental delegando decisiones a los atajos heurísticos. Aunque el Sistema 2 posee la capacidad de supervisar y corregir respuestas intuitivas, en numerosas ocasiones no logra intervenir adecuadamente, permitiendo que emociones, impulsos y mecanismos heurísticos condicionen la toma de decisiones.

En mercados caracterizados por incertidumbre y volatilidad, las decisiones suelen verse influenciadas por respuestas rápidas asociadas al Sistema 1, especialmente durante períodos de estrés financiero, lo que explica fenómenos como el pánico de mercado, la euforia, las sobrerreacciones frente a noticias y el comportamiento de manada. La comprensión de estos mecanismos constituye la base conceptual para analizar los principales sesgos cognitivos y emocionales que afectan la gestión activa humana, cuestión que será desarrollada en el apartado siguiente.

2.1.3 Sesgos cognitivos y emocionales aplicados a inversiones

Según Shefrin (2000), las finanzas conductuales reconocen que los inversores utilizan reglas generales o atajos mentales para procesar información y tomar decisiones bajo incertidumbre.

En este sentido, numerosos comportamientos observados en los mercados financieros pueden explicarse a partir de sesgos cognitivos y emocionales que alteran la percepción del riesgo, la interpretación de información y la administración de carteras.

Representatividad (Representativeness Bias)

El sesgo de representatividad, desarrollado por Tversky y Kahneman (1974), hace referencia a la tendencia de los individuos a realizar predicciones basándose en patrones o similitudes observadas previamente, ignorando muchas veces probabilidades objetivas o información estadística relevante. Dentro de los mercados financieros, este sesgo puede llevar a extrapolar tendencias recientes hacia el futuro, asumiendo que

determinados activos continuarán comportándose de la misma manera únicamente porque así ocurrió en períodos anteriores.

Aversión a la pérdida (Loss Aversion)

Kahneman y Tversky (1979) sostienen que las pérdidas generan un impacto emocional significativamente mayor que las ganancias equivalentes. Dentro del proceso de inversión, este sesgo suele generar comportamientos irracionales como mantener posiciones perdedoras durante períodos prolongados o evitar decisiones de inversión por miedo a pérdidas futuras. Según Shefrin y Statman (1985), este comportamiento se encuentra estrechamente relacionado con el disposition effect, donde los inversores venden activos ganadores demasiado rápido mientras mantienen activos perdedores esperando recuperar pérdidas.

Comportamiento de manada (Herding Behavior)

El comportamiento de manada hace referencia a la tendencia de los individuos a imitar decisiones de otros participantes del mercado aun cuando no existan fundamentos propios suficientes que justifiquen dicha conducta. Según Bikhchandani y Sharma (2000), este fenómeno suele intensificarse durante períodos de incertidumbre y elevada volatilidad financiera. En los mercados financieros, el comportamiento de manada puede generar movimientos excesivos de precios, burbujas especulativas y episodios de pánico financiero.

Disponibilidad (Availability Bias)

El sesgo de disponibilidad, desarrollado por Kahneman y Tversky (1973), describe la tendencia de los individuos a otorgar mayor importancia a información reciente, fácilmente recordable o emocionalmente impactante. En consecuencia, los inversores suelen sobreestimar la relevancia de noticias recientes o eventos de mercado que reciben mayor cobertura mediática.

Confirmación (Confirmation Bias)

El sesgo de confirmación hace referencia a la tendencia de los individuos a buscar información que valide sus creencias previas mientras ignoran evidencia contradictoria. Según Nickerson (1998), este mecanismo afecta significativamente la capacidad de tomar decisiones objetivas y racionales.

En mercados financieros, este sesgo puede llevar a que los inversores mantengan posiciones de inversión aun cuando exista información negativa relevante, simplemente porque priorizan noticias o análisis compatibles con sus expectativas iniciales.

Efecto dotación (Endowment Effect)

El efecto dotación, desarrollado por Thaler (1980), sostiene que los individuos tienden a sobrevalorar aquellos activos que ya poseen simplemente por el hecho de pertenecerles. Dentro de los mercados financieros, este sesgo puede generar un apego emocional excesivo hacia determinadas inversiones, dificultando decisiones racionales de venta o rebalanceo de cartera.

Falacia del jugador (Gambler's Fallacy)

La falacia del jugador, también conocida como falacia de Montecarlo, hace referencia a la creencia errónea de que eventos aleatorios futuros pueden verse influenciados por secuencias de eventos pasados, aun cuando estos sean estadísticamente independientes entre sí. Según Shefrin (2000), este sesgo surge cuando los individuos asumen incorrectamente que determinados resultados "deberían compensarse" en el corto plazo. Dentro de los mercados financieros, este comportamiento puede observarse cuando los inversores consideran que un activo necesariamente debe recuperarse luego de varias caídas consecutivas o, por el contrario, que una tendencia alcista se encuentra próxima a revertirse simplemente por haber persistido durante mucho tiempo.

Sesgo de atribución personal (Self-Attribution Bias)

El sesgo de atribución personal hace referencia a la tendencia de los individuos a atribuir los resultados positivos a sus propias habilidades mientras responsabilizan a factores externos por los resultados negativos. Según Shefrin (2000), este mecanismo psicológico contribuye a reforzar percepciones exageradas de capacidad y control dentro del proceso de inversión. En el ámbito financiero, este sesgo puede llevar a que inversores y gestores profesionales sobreestimen su habilidad para seleccionar activos o anticipar movimientos de mercado luego de períodos de buenos resultados.

Sesgo de actualidad y regla del final de pico (Recency Bias and Peak-End Rule)

El sesgo de actualidad describe la tendencia de los individuos a otorgar mayor relevancia a eventos recientes al momento de proyectar escenarios futuros. Complementariamente, Kahneman (2011) sostiene mediante la regla del final de pico (Peak-End Rule) que las

personas suelen evaluar experiencias principalmente en función de los momentos más intensos y del resultado final, antes que considerando el proceso completo. Dentro de los mercados financieros, estos mecanismos pueden generar sobrerreacciones frente a eventos recientes de mercado, llevando a los inversores a extrapolar comportamientos de corto plazo hacia el futuro.

2.1.4 Overconfidence en gestores profesionales

Kahneman (2011) sostiene que las personas tienden a sobreestimar el grado en que comprenden el mundo y la precisión de sus propios juicios. El exceso de confianza (overconfidence bias) es uno de los sesgos más estudiados dentro de las finanzas conductuales, definido como la tendencia de los individuos a sobreestimar sus conocimientos, habilidades y capacidad para interpretar información financiera, y documentado tanto en inversores individuales como en gestores profesionales. Puetz y Ruenzi (2011) concluyeron que muchos profesionales presentan niveles significativos de este sesgo, atribuyendo los buenos resultados a su propia habilidad mientras explican los negativos mediante factores externos. Barber y Odean (2001) sostienen que este exceso de confianza lleva a operar con mayor frecuencia por la creencia de poseer información superior, lo que según Puetz y Ruenzi (2011) deteriora el rendimiento ajustado por riesgo de los fondos administrados: los gestores excesivamente confiados asumen posiciones más agresivas, reaccionan de manera impulsiva frente a cambios de mercado y subestiman escenarios adversos.

2.1.5 ¿Skill o suerte? La performance de la gestión activa

Kahneman (2011) sostiene que las personas tienden a atribuir sentido y causalidad a eventos que muchas veces pueden responder al azar. Uno de los principales debates dentro de la teoría financiera se centra en determinar si los gestores activos poseen realmente la capacidad de generar rendimientos superiores de manera consistente o si gran parte de dichos resultados responden simplemente al azar. Fama y French (2010) analizaron la persistencia de performance en fondos de inversión y concluyeron que existe escasa evidencia de habilidad sistemática (skill) dentro de la gestión activa, especialmente luego de considerar costos operativos y comisiones, ya que una parte significativa de la performance observada puede explicarse por factores aleatorios antes que por capacidades superiores de selección de activos.

En la misma línea, Puetz y Ruenzi (2011) demostraron que muchos administradores de fondos incrementan significativamente sus niveles de trading luego de obtener buenos resultados previos, comportamiento asociado al exceso de confianza, atribuyendo los buenos resultados a su propia habilidad y subestimando el papel del azar. Barber y Odean (2001) sostienen que este mayor nivel de trading aumenta los costos de transacción y deteriora el rendimiento ajustado por riesgo de las carteras administradas, un efecto particularmente problemático en contextos de elevada volatilidad e incertidumbre financiera.

En consecuencia, la evidencia desarrollada tanto por la teoría financiera como por las finanzas conductuales sugiere que la generación de alpha persistente dentro de la gestión activa representa un desafío considerable. Dentro del presente trabajo, estas cuestiones adquieren especial importancia debido a que los fondos comunes de inversión de renta variable argentina serán utilizados como referencia para comparar el desempeño de estrategias sistemáticas construidas mediante modelos de machine learning.

3. La gestión algorítmica: machine learning aplicado a la predicción de retornos

3.1 Aprendizaje supervisado en finanzas: panorama general

El aprendizaje supervisado es la rama del machine learning en la que un algoritmo aprende a estimar una función que vincula un conjunto de variables de entrada (features) con una variable de salida (target) a partir de un conjunto de ejemplos etiquetados. En el contexto de la predicción de retornos financieros, las variables de entrada típicamente incluyen información de precios pasados, indicadores de momentum, medidas de volatilidad y otras características construidas a partir de datos históricos, mientras que la variable de salida es el retorno futuro del activo, ya sea como valor numérico (problema de regresión) o como categoría binaria que indica si el activo subirá o bajará (problema de clasificación). Esta segunda formulación es la que se adopta en la presente investigación.

Dentro de esta literatura, el trabajo de Gu et al. (2020), publicado en The Review of Financial Studies bajo el título Empirical Asset Pricing via Machine Learning, constituye una de las referencias más exhaustivas y citadas. Los autores se proponen explícitamente *"sintetizar el campo del machine learning con el problema canónico del asset pricing empírico: la medición de las primas de riesgo de los activos"* (Gu et al.,

2020, p. 2223), realizando un análisis comparativo de un amplio repertorio de métodos que incluye modelos lineales generalizados, técnicas de reducción de dimensionalidad, árboles de regresión potenciados (boosted regression trees), random forests y redes neuronales aplicados a la predicción de retornos en el mercado accionario estadounidense. Una de las conclusiones centrales es que el machine learning ofrece una descripción mejorada del comportamiento de los retornos esperados frente a los métodos tradicionales, atribuyendo sus ganancias predictivas a *"la incorporación de interacciones no lineales entre predictores que otros métodos no capturan"* (Gu et al., 2020, p. 2224). Esta capacidad es precisamente la que diferencia a los modelos de machine learning de los enfoques tradicionales basados en regresión lineal, y la que justifica conceptualmente su elección para la presente investigación, ya que los retornos financieros suelen estar gobernados por dinámicas complejas que rara vez pueden capturarse con relaciones lineales simples. Asimismo, los autores señalan que *"todos los métodos coinciden en el mismo conjunto reducido de señales predictivas dominantes que incluye variaciones de momentum, liquidez y volatilidad"* (Gu et al., 2020, p. 2223), evidencia particularmente relevante dado que las variables predictivas utilizadas en este trabajo son retornos pasados, momentum a 5 y 20 días, y volatilidad a 20 días se ubican exactamente dentro de ese conjunto.

El trabajo de Krauss et al. (2017), publicado en el European Journal of Operational Research bajo el título Deep Neural Networks, Gradient-Boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500, ofrece un complemento fundamental al concentrarse específicamente en las tres familias de modelos que constituyen, junto con Markowitz (1952), el núcleo metodológico del presente TIF. Los autores implementan redes neuronales profundas, gradient-boosted trees y random forests sobre las acciones del S&P 500 durante el período 1992–2015, prediciendo la probabilidad de que cada acción supere al mercado en el día siguiente y construyendo carteras a partir de las posiciones extremas de ese ranking. Esta formulación como clasificación binaria comparte rasgos directos con la metodología adoptada en este trabajo, en el que cada modelo estima la probabilidad de que una acción suba o baje en el día T+1 a partir de la información disponible hasta el día T, lo que permite tomar a Krauss et al. (2017) no solo como antecedente de validación general sino como punto de comparación cercano a nivel de diseño experimental. Los autores concluyen que sus hallazgos *"plantean un desafío severo a la forma semi-fuerte de la hipótesis de mercados eficientes"* (Krauss et

al., 2017, p. 689), al demostrar que es posible generar retornos sistemáticamente superiores al benchmark utilizando exclusivamente información públicamente disponible.

Sin embargo, ambos estudios comparten una limitación importante para los fines de esta investigación: se concentran en mercados desarrollados, específicamente el mercado accionario estadounidense, con universos de activos caracterizados por elevada liquidez y alta cobertura analítica. La extensión de estos resultados al mercado argentino, con sus particularidades institucionales, su menor profundidad y sus episodios recurrentes de estrés financiero, constituye una pregunta empírica abierta que el presente trabajo se propone abordar.

3.2 Random Forest

El algoritmo Random Forest, propuesto por Leo Breiman en su artículo Random Forests publicado en la revista Machine Learning en 2001, constituye uno de los métodos de aprendizaje supervisado más utilizados tanto en investigación académica como en aplicaciones prácticas. Su éxito se explica por una combinación de simplicidad conceptual, robustez frente a datos ruidosos y, particularmente para los fines de este trabajo, su capacidad para capturar relaciones no lineales entre variables sin requerir supuestos paramétricos sobre la distribución de los datos.

La unidad básica de construcción de un Random Forest es el árbol de decisión, una estructura que clasifica observaciones aplicando, en cada nodo, una regla de partición sobre alguna de las variables predictivas hasta arribar a una clasificación final en los nodos terminales. El árbol individual presenta dos limitaciones importantes: tiende a sobre ajustarse a los datos de entrenamiento cuando se le permite crecer sin restricciones y es altamente sensible a pequeñas perturbaciones en la muestra. Estas limitaciones motivaron el desarrollo de métodos ensemble que combinan las predicciones de muchos árboles para obtener un clasificador más estable y preciso (Breiman, 2001). La definición formal del algoritmo es expresada por Breiman (2001) en términos directos*: "un random forest es un clasificador consistente en una colección de clasificadores estructurados como árboles (...), y cada árbol emite un voto unitario por la clase más probable para cada entrada x"* (Breiman, 2001, p. 6).

La contribución central de Breiman (2001) consiste en la introducción simultánea de dos fuentes de aleatoriedad durante la construcción de cada árbol. La primera es el bagging (bootstrap aggregating), técnica desarrollada previamente por el propio Breiman (1996),

por la cual cada árbol se entrena sobre una muestra obtenida con reemplazo del conjunto original, dejando aproximadamente un tercio de las observaciones fuera de cada muestra el conjunto out-of-bag que permite estimar el error de generalización sin necesidad de un set de validación separado. La segunda, y la innovación distintiva de Random Forest, es la selección aleatoria de variables en cada partición: en lugar de evaluar todas las variables disponibles, el algoritmo selecciona aleatoriamente un subconjunto reducido de candidatas y elige la que produce la mejor separación. Al limitar la influencia de las variables más fuertes, fuerza a los distintos árboles a explorar patrones diferentes, reduciendo la correlación entre sus predicciones y mejorando la precisión del conjunto. Estas dos fuentes de aleatoriedad, combinadas, confieren al método sus propiedades favorables de robustez y generalización (Breiman, 2001).

En el contexto de la presente investigación, Random Forest se utiliza como uno de los tres modelos predictivos cuya salida alimenta el problema de optimización de Markowitz (1952) formulado en la sección 1.1. Su rol consiste en estimar, para cada acción del panel y para cada jornada de la simulación, la probabilidad de que el retorno del día siguiente sea positivo. Esta aplicación se inscribe dentro de las prácticas documentadas en la literatura empírica revisada en la sección 3.1, particularmente en Krauss et al. (2017), donde Random Forest aparece como uno de los tres modelos sistemáticamente evaluados en el contexto de arbitraje estadístico. Los detalles específicos de la configuración utilizada se desarrollan en el capítulo de metodología.

3.3 Gradient boosting: XGBoost

El segundo modelo del núcleo metodológico de esta investigación es XGBoost (eXtreme Gradient Boosting), una implementación escalable del algoritmo de gradient boosting desarrollada por Tianqi Chen y Carlos Guestrin y presentada en el artículo XGBoost: A Scalable Tree Boosting System, publicado en las actas de la conferencia KDD en 2016. A diferencia de Random Forest, que combina árboles entrenados en paralelo e independientemente, el gradient boosting construye los árboles secuencialmente: cada nuevo árbol se entrena para corregir los errores cometidos por el conjunto de árboles previos, de modo que el modelo final es una suma ordenada en la que cada árbol aporta una corrección marginal sobre los anteriores (Chen & Guestrin, 2016). Esta lógica permite alcanzar una mayor precisión predictiva, aunque al costo de un mayor riesgo de sobreajuste si los hiperparámetros no se controlan adecuadamente.

Una característica especialmente relevante para los fines de este trabajo es la regularización. Mientras que Random Forest controla el sobreajuste mediante la aleatoriedad introducida en bagging y selección de variables, XGBoost lo controla mediante penalizaciones directas a la complejidad de cada árbol: durante el entrenamiento, penaliza a los árboles que crecen demasiado o que asignan pesos muy grandes a sus hojas, forzando al modelo a mantener una estructura más simple (Chen & Guestrin, 2016). Este balance permite construir modelos altamente expresivos sin sacrificar capacidad de generalización, una propiedad particularmente valiosa cuando los datos disponibles son ruidosos o limitados, como suele ocurrir en mercados emergentes.

En el contexto de la literatura financiera revisada en la sección 3.1, XGBoost aparece como uno de los modelos sistemáticamente evaluados tanto por Gu et al. (2020) como por Krauss et al. (2017), donde es referido como gradient-boosted trees (GBT), figurando en ambos trabajos entre los métodos de mejor desempeño predictivo. En la implementación de la presente investigación, se entrena un clasificador binario por acción que estima, día a día, la probabilidad de que el retorno del día siguiente sea positivo, siguiendo las recomendaciones de regularización y configuración documentadas por Chen y Guestrin (2016). Los detalles se desarrollan en el capítulo de metodología.

3.4 Redes neuronales recurrentes: LSTM

El tercer modelo del núcleo metodológico de esta investigación es la red neuronal recurrente con arquitectura LSTM (Long Short-Term Memory), propuesta por Sepp Hochreiter y Jürgen Schmidhuber en su artículo Long Short-Term Memory, publicado en la revista Neural Computation en 1997. A diferencia de Random Forest y XGBoost, que tratan cada observación como independiente, las redes LSTM están diseñadas específicamente para procesar secuencias temporales: aprenden de la información que llega en orden, donde el resultado predicho en cada momento puede depender de lo ocurrido en momentos previos (Hochreiter & Schmidhuber, 1997). La contribución central de los autores consiste en la introducción de una celda de memoria protegida por compuertas que regulan, en cada paso, qué información nueva entra, qué información previa se conserva y qué información se utiliza como salida, permitiendo a la red "recordar" patrones relevantes del pasado lejano sin que se diluyan mientras descarta lo que ha dejado de ser útil (Hochreiter & Schmidhuber, 1997).

En el contexto de la predicción de retornos financieros, las redes LSTM resultan particularmente atractivas por dos razones. Primero, los precios y retornos de los activos son datos secuenciales con dependencias temporales: el comportamiento de un activo en un día determinado no es independiente de lo ocurrido en los días previos. Segundo, las LSTM capturan relaciones no lineales entre variables, propiedad compartida con Random Forest y XGBoost pero ejercida sobre la dimensión temporal. Tanto Krauss et al. (2017) como Gu et al. (2020) incluyen redes neuronales entre los modelos evaluados, con resultados que respaldan empíricamente su uso en problemas de predicción de retornos.

La elección de LSTM como tercer modelo responde a un criterio de diversificación de enfoques: mientras que Random Forest y XGBoost son métodos basados en árboles que comparten un mismo paradigma de aprendizaje, las redes LSTM operan sobre fundamentos completamente distintos mediante procesamiento secuencial con memoria interna, representando una familia de modelos cuyo desempeño relativo en mercados emergentes constituye en sí mismo una pregunta empírica abierta. En la implementación de este trabajo, las redes LSTM resuelven el mismo problema de clasificación binaria que los modelos anteriores, estimando para cada acción y cada día la probabilidad de que el retorno del día siguiente sea positivo, con mecanismos de regularización para evitar el sobreajuste. Las especificaciones técnicas se desarrollan en el capítulo de metodología.

3.5 Variables predictivas: momentum y volatilidad

El desempeño de los modelos de aprendizaje supervisado depende críticamente de las variables que se utilicen como insumo. En la presente investigación, los tres modelos se alimentan de un conjunto reducido pero teóricamente fundamentado de variables: indicadores de momentum y de volatilidad, construidos a partir de los precios históricos de cada acción.

El efecto momentum, la tendencia de las acciones con retornos elevados recientes a continuar superando al mercado en los meses siguientes, fue documentado por Jegadeesh y Titman (1993) en su artículo Returns to Buying Winners and Selling Losers, publicado en The Journal of Finance. Los autores demuestran que estrategias que compran las acciones con mejor desempeño reciente y venden las de peor desempeño generan retornos significativamente positivos durante los 3 a 12 meses siguientes (Jegadeesh & Titman, 1993), hallazgo que desafió directamente la forma débil de la

hipótesis de mercados eficientes revisada en la sección 1.2. La validez del momentum como predictor ha sido confirmada por evidencia empírica posterior, incluyendo los trabajos de Gu et al. (2020) y Krauss et al. (2017), donde sus variantes aparecen entre las señales más informativas. La volatilidad histórica, calculada como desviación estándar de los retornos sobre una ventana móvil, constituye la segunda variable predictiva relevante, sustentada en el clustering de volatilidad y en su asociación con la prima de riesgo del activo. La especificación exacta de las ventanas temporales y los detalles de construcción de ambas variables se desarrollan en el capítulo de metodología.

3.6 Sesgos y riesgos metodológicos de los modelos algorítmicos

A diferencia de los sesgos conductuales analizados en el capítulo anterior, los sesgos algorítmicos no provienen de emociones o errores psicológicos individuales, sino de limitaciones asociadas a los datos, al diseño del modelo y al contexto en el cual el algoritmo es aplicado. Mehrabi et al. (2021) señalan que los sesgos en sistemas de machine learning pueden originarse tanto en los datos utilizados para entrenar el modelo como en las decisiones técnicas adoptadas durante su diseño e implementación. Esta distinción resulta fundamental para evaluar de manera equilibrada las limitaciones de ambos enfoques de gestión: los humanos enfrentan sesgos conductuales; los algoritmos, sesgos estadísticos, técnicos y metodológicos. A continuación se desarrollan los más relevantes para el diseño y la evaluación de la presente investigación.

Sesgo de datos (Data Bias)

El sesgo de datos aparece cuando la información utilizada para entrenar un modelo no representa adecuadamente el fenómeno que se busca predecir (Mehrabi et al., 2021). En mercados emergentes como el argentino, la baja liquidez, la menor profundidad del mercado y la presencia de episodios extremos pueden hacer que los datos disponibles reflejen condiciones particulares y no necesariamente patrones estables, de modo que el modelo no aprende una verdad permanente del mercado sino las regularidades presentes en la muestra histórica utilizada.

Look-ahead bias

Este sesgo ocurre cuando un modelo utiliza información que no habría estado disponible al momento real de tomar la decisión de inversión, generando resultados artificialmente superiores en el backtest. López de Prado (2018) enfatiza la importancia de respetar la

estructura temporal de los datos para evitar errores de validación que inflen artificialmente los resultados. En esta investigación, el riesgo se controla explícitamente: los modelos se entrenan únicamente con información previa al inicio de la simulación, sin reentrenamiento posterior, y las predicciones se generan utilizando exclusivamente datos disponibles hasta el día anterior.

Sobreajuste y backtest overfitting

El sobreajuste fue definido y formalizado por Bailey et al. (2015) en su artículo The Probability of Backtest Overfitting. El fenómeno consiste en seleccionar como "buenas" estrategias que solo son aparentemente exitosas porque sus parámetros fueron ajustados a las particularidades aleatorias del histórico, sin que esa performance se replique en datos futuros. Los autores advierten que *"las técnicas estadísticas estándar diseñadas para prevenir el sobreajuste, como el método hold-out, tienden a ser poco confiables e imprecisas en el contexto de backtests de inversión"* (Bailey et al., 2015, p. 39). López de Prado (2018) amplía esta advertencia señalando que repetir experimentos sobre los mismos datos, comunicar solo resultados favorables o ajustar hiperparámetros mirando el set de prueba constituyen formas encubiertas de overfitting que invalidan las conclusiones empíricas.

Sesgo por cambio de régimen (Dataset Shift)

El dataset shift aparece cuando la distribución de los datos de entrenamiento difiere de la distribución de los datos de aplicación. Quiñonero-Candela et al. (2009) explican que este problema afecta la capacidad de generalización del modelo cuando las condiciones del conjunto de entrenamiento y del conjunto de prueba no son equivalentes. Este sesgo es especialmente relevante para esta investigación dado que el período 2019–2026 incluye episodios de estrés financiero de naturaleza muy diferente entre sí, detallados en la sección 6.2. Justamente por eso, estos períodos funcionan como prueba de robustez, transformando una limitación potencial en una oportunidad analítica.

Aversión al algoritmo (Algorithm Aversion)

La aversión al algoritmo, documentada por Dietvorst, Simmons y Massey (2015), hace referencia a la tendencia de los individuos a confiar menos en las predicciones de un sistema algorítmico que en las de un humano, incluso cuando el algoritmo demuestra un desempeño superior. Según los autores, este rechazo se intensifica una vez que las personas observan al algoritmo cometer un error. Dentro de los mercados financieros,

este sesgo puede explicar por qué los gestores profesionales tienden a preferir su propio juicio discrecional por sobre modelos sistemáticos, incluso ante evidencia de desempeño más consistente.

Sesgo de liquidez y ejecución (Liquidity Bias)

El sesgo de liquidez y ejecución aparece cuando una estrategia muestra buenos resultados teóricos pero enfrenta dificultades para implementarse en condiciones reales de mercado. Este sesgo es especialmente relevante en el mercado argentino, dado que la menor liquidez y profundidad relativa constituyen características estructurales de los mercados emergentes (Bekaert & Harvey, 2002). En esta investigación, una diferencia entre el desempeño simulado y el ejecutable podría surgir no por error del modelo sino por fricciones propias del mercado local, lo que acota el alcance interpretativo de los resultados: lo que se evalúa es el desempeño teórico bajo condiciones idealizadas, no la rentabilidad efectiva en operación real.

Con la presente sección se cierra el capítulo de gestión algorítmica: los modelos de machine learning ofrecen capacidad predictiva genuina respaldada por la literatura empírica revisada en las secciones anteriores, pero esa capacidad solo puede traducirse en hallazgos válidos si el diseño experimental controla rigurosamente las fuentes de sobreajuste y reconoce las limitaciones contextuales del mercado donde se aplica. El capítulo siguiente desplaza el foco del aparato técnico a su contexto de aplicación: las características del mercado argentino y los períodos de estrés financiero que estructuran la simulación.

4. El terreno de la comparación: mercados emergentes y el mercado argentino

4.1 Mercados emergentes: predictibilidad e ineficiencias estructurales

Según Bekaert y Harvey (2002), los mercados emergentes se caracterizan por presentar menores niveles de liquidez, menor profundidad de mercado y una mayor sensibilidad frente a shocks económicos, políticos y financieros en comparación con los mercados desarrollados.

A diferencia de los mercados desarrollados, donde la información tiende a incorporarse rápidamente a los precios de los activos, los mercados emergentes presentan mayores asimetrías informacionales, menor cobertura analítica y una participación más limitada de inversores institucionales. Como consecuencia, variables vinculadas al contexto político, cambiario y macroeconómico adquieren una influencia significativa sobre la dinámica del mercado financiero, incrementando los niveles de volatilidad y ruido dentro del proceso de valuación de activos.

En este sentido, Harvey (1995) sostiene que los mercados emergentes presentan mayores niveles de predictibilidad relativa respecto de mercados desarrollados debido a menores niveles de integración y eficiencia financiera. Estas características adquieren especial relevancia dentro del presente trabajo, dado que permiten justificar la utilización del mercado argentino como escenario de análisis para comparar el desempeño de estrategias de gestión humana y modelos sistemáticos de inversión basados en machine learning.

4.2 El mercado argentino: BYMA, controles cambiarios y el dólar CCL como denominador

El mercado financiero argentino se caracteriza históricamente por presentar elevados niveles de volatilidad macroeconómica, recurrentes restricciones cambiarias y una fuerte sensibilidad frente a eventos políticos y económicos. A diferencia de mercados desarrollados, donde las condiciones monetarias y cambiarias suelen presentar mayor estabilidad, el mercado argentino opera frecuentemente bajo escenarios de inflación elevada, depreciación monetaria y cambios regulatorios significativos.

Dentro de este contexto, Bolsas y Mercados Argentinos (BYMA) constituye el principal ámbito de negociación de activos financieros del país, concentrando gran parte de la

operatoria de renta variable y otros instrumentos vinculados al mercado de capitales local. Sin embargo, el mercado argentino presenta menor profundidad y liquidez en comparación con mercados internacionales desarrollados, situación que incrementa la sensibilidad de los precios frente a shocks externos, eventos políticos y cambios en las expectativas macroeconómicas.

Asimismo, uno de los principales factores que condiciona el funcionamiento del mercado financiero argentino es la presencia recurrente de controles cambiarios y múltiples tipos de cambio. Estas restricciones generan distorsiones relevantes dentro del proceso de valuación de activos financieros, dificultando la comparación de rendimientos medidos en moneda local. En este escenario, el dólar contado con liquidación (CCL) adquiere especial relevancia como referencia financiera para expresar retornos en moneda dura y reducir el efecto distorsivo de la inflación y la depreciación del peso argentino.

El dólar CCL surge mediante operaciones de compra y venta de activos financieros entre mercados locales e internacionales, permitiendo obtener divisas fuera del mercado oficial de cambios. Debido a su utilización como referencia financiera dentro del mercado argentino, el CCL se consolidó como uno de los principales indicadores utilizados por inversores institucionales y participantes del mercado para evaluar cobertura cambiaria, valuación de activos y rendimientos reales de inversión.

Dentro del presente trabajo, la utilización del dólar CCL como denominador común permite homogeneizar los rendimientos de las estrategias analizadas y capturar de manera más precisa el riesgo financiero asociado al mercado argentino.

4.3 Períodos de estrés financiero como contextos críticos de evaluación

Según Bekaert y Harvey (2002), los períodos de inestabilidad financiera permiten identificar con mayor claridad las vulnerabilidades y limitaciones presentes en los mercados emergentes. Los períodos de estrés financiero constituyen escenarios particularmente relevantes para evaluar estrategias de inversión dado que incrementan los niveles de volatilidad, incertidumbre y sensibilidad frente a nueva información, generando cambios abruptos en las expectativas y mayor dispersión de precios. En este trabajo, la selección de cuatro períodos de estrés responde a la necesidad de evaluar el comportamiento relativo entre ambos enfoques de gestión bajo contextos de elevada incertidumbre política, económica y cambiaria.

Los cuatro episodios seleccionados cubren un espectro amplio de tipos de estrés: las PASO 2019, caracterizadas por una fuerte corrección tras el resultado electoral que evidenció la elevada sensibilidad de los activos locales frente a eventos políticos; el COVID-19 en 2020, uno de los episodios de mayor volatilidad financiera global de las últimas décadas; el período de 2023, donde a pesar de la elevada inflación e incertidumbre macroeconómica los activos registraron un fuerte rally impulsado por expectativas de cambio de régimen; y el escenario electoral de 2025, caracterizado por reconfiguración de expectativas económicas y elevada sensibilidad política. Su selección responde a su capacidad para funcionar como escenarios de prueba frente a contextos extremos de mercado, permitiendo evaluar la robustez de las estrategias analizadas.

5. Antecedentes empíricos: comparaciones humano vs. máquina y el vacío en mercados emergentes

Los capítulos anteriores presentaron, por separado, las características y limitaciones de la gestión humana, los fundamentos técnicos del aprendizaje automatizado aplicado a finanzas y las particularidades del mercado argentino donde se desarrollará la simulación. Este capítulo final reúne ambas dimensiones revisando la evidencia empírica disponible sobre comparaciones directas entre gestión humana y algorítmica de carteras. Dos trabajos recientes ,ambos centrados en la industria de hedge funds en mercados desarrollados ,ofrecen los antecedentes más cercanos a la pregunta de investigación de este TIF, y permiten identificar el vacío específico que el presente trabajo se propone llenar.

5.1 Comparaciones discrecional vs. sistemático en hedge funds

El estudio de Harvey et al. (2017), titulado Man vs. Machine: Comparing Discretionary and Systematic Hedge Fund Performance y publicado en The Journal of Portfolio Management, constituye uno de los antecedentes empíricos más relevantes para la pregunta planteada en esta investigación. Los autores comparan el desempeño de fondos discrecionales y sistemáticos durante el período 1996–2014 sobre una muestra de más de 9.000 fondos de la base Hedge Fund Research (HFR).

La motivación del trabajo nace de una observación curiosa: a pesar de la sofisticación creciente de los enfoques sistemáticos, una proporción significativa de los grandes asignadores de capital evita destinar fondos a estrategias automatizadas. Harvey et al. (2017) atribuyen esta reticencia a un fenómeno conocido como aversión al algoritmo

(algorithm aversion), por el cual los seres humanos tienden a desconfiar más de las decisiones automatizadas que de las humanas, incluso cuando las primeras presentan mejor desempeño. Los autores afirman, sin embargo, que *"no encontramos base empírica para tal aversión"* (Harvey et al., 2017, p. 55).

Tras ajustar los retornos por factores de riesgo tradicionales, dinámicos y de volatilidad, los autores concluyen que *"el desempeño de los gestores sistemáticos y discrecionales es similar, una vez ajustado por volatilidad y exposiciones a factores"* (Harvey et al., 2017, p. 55), con leves ventajas para los sistemáticos en categorías como macro y equity. Un hallazgo adicional resulta relevante: el desempeño de los fondos discrecionales se explica en mayor proporción por exposiciones a factores de riesgo conocidos, lo que sugiere que son los gestores humanos quienes obtienen una porción mayor de sus retornos a partir de exposiciones replicables a factores estándar. La principal limitación del trabajo para los fines de esta tesis es su foco geográfico: la totalidad de la muestra proviene de mercados desarrollados, dejando abierta la pregunta de cómo se comportaría dicha comparación en un contexto emergente como el argentino.

5.2 Inteligencia artificial en hedge funds

El trabajo de Grobys et al. (2022), titulado Man Versus Machine: On Artificial Intelligence and Hedge Funds Performance y publicado en Applied Economics, complementa la línea iniciada por Harvey et al. (2017) introduciendo una clasificación en cuatro niveles de automatización. Grobys et al. (2022) encuentran que los fondos con mayor nivel de automatización superan en desempeño a los de mayor intervención humana: una estrategia de costo cero larga en los más automatizados y corta en los de mayor intervención genera un spread de al menos 50 puntos básicos por mes, concluyendo que la automatización juega un rol importante en la rentabilidad de la industria de hedge funds (Grobys et al., 2022).

Esta investigación presenta dos limitaciones relevantes para los fines de este TIF: la muestra proviene íntegramente de mercados desarrollados, y ambos trabajos se concentran en hedge funds vehículos profesionalizados, escasamente regulados y restringidos a inversores calificados mientras que la presente investigación toma como referencia los fondos comunes de inversión de renta variable, vehículos retail accesibles al público general y sometidos a regulación más estricta.

5.3 El vacío de literatura: ausencia de evidencia comparativa en el mercado argentino

La revisión de los antecedentes deja al descubierto una carencia de investigación que constituye una de las principales justificaciones del presente trabajo. Tanto Harvey et al. (2017) como Grobys et al. (2022) abordan la comparación entre gestión humana y algorítmica en mercados desarrollados, dejando inexplorada la pregunta de cómo se comportaría dicha comparación en un mercado emergente con las particularidades del argentino. Esta limitación geográfica resulta especialmente relevante a la luz del análisis de Bekaert y Harvey (2002): si los modelos algorítmicos demostraron capacidad para superar al menos parcialmente a la gestión humana en mercados altamente eficientes, cabe preguntarse si esa ventaja será aún mayor en contextos donde el margen de explotación de ineficiencias es estructuralmente más amplio. A esto se suma que la literatura disponible se concentra en hedge funds, mientras que el presente trabajo utiliza como benchmark los fondos comunes de inversión de renta variable argentina vehículos retail que constituyen el principal canal por el cual el inversor minorista local accede a una cartera diversificada y no se identificaron estudios comparables que apliquen modelos de machine learning al universo de acciones argentinas.

En conjunto, tres elementos configuran el vacío específico que esta investigación se propone abordar: la ausencia de evidencia comparativa en mercados emergentes, la ausencia de comparaciones que utilicen FCI minoristas como benchmark humano, y la ausencia de aplicaciones de machine learning al mercado argentino. Los hallazgos del trabajo, cualquiera sea su signo, contribuirán a llenar parcialmente este espacio en la literatura aportando evidencia empírica desde un contexto poco estudiado.

6. Metodología

6.1 Enfoque metodológico y diseño de la investigación

La presente investigación adopta un enfoque cuantitativo, dado que se basa en el análisis de series de datos financieros y en la medición de variables objetivas como rentabilidad, volatilidad y riesgo. En línea con lo señalado por Sautu (2010), la metodología construye un conjunto de procedimientos orientados a producir evidencia empírica articulada con los objetivos de la investigación, lo que en este caso implica diseñar una simulación replicable que permita comparar el desempeño de distintas estrategias de gestión de carteras bajo condiciones históricas reales.

El diseño es no experimental y longitudinal. Es no experimental porque no se manipulan variables ni se interviene sobre el mercado, los datos utilizados corresponden a precios y retornos históricos observados, sobre los cuales se aplican los modelos sin alterar las condiciones del fenómeno estudiado. Es longitudinal porque el análisis abarca un período extendido de tiempo de 2019 a 2026, que incluye distintos regímenes de mercado, lo que permite evaluar el comportamiento de las estrategias bajo condiciones cambiantes y no solo en un momento puntual.

En cuanto a su alcance, la investigación es de tipo comparativa y explicativa. Es comparativa porque su objetivo central consiste en contrastar el desempeño de carteras gestionadas mediante modelos de Machine Learning frente a carteras administradas por gestores humanos, medido a través de métricas estandarizadas de rendimiento y riesgo. Es explicativa porque no se limita a describir las diferencias observadas, sino que busca interpretarlas a la luz del marco teórico desarrollado, identificando en qué contextos cada enfoque presenta mejores resultados y qué factores estructurales, sesgos conductuales, costos operativos o capacidad de procesamiento de información pueden dar cuenta de esas diferencias.

Este diseño responde directamente a la pregunta de investigación central del trabajo: para determinar si los modelos de Machine Learning superan a la gestión humana durante episodios de estrés financiero, es necesario construir una simulación que reproduzca fielmente las condiciones de mercado históricas, aplique los modelos con información disponible en cada momento sin filtración de datos futuros, y compare los resultados contra un benchmark humano representativo del universo de gestión activa local.

6.2 Período de análisis y justificación

El período de análisis comprende desde el 3 de junio de 2019 hasta febrero de 2026, abarcando aproximadamente seis años y medio de operatoria continua en el mercado argentino. La elección de este período responde a la necesidad de cubrir distintos regímenes de mercado y episodios de estrés financiero de naturaleza heterogénea, que permitan evaluar la robustez de las estrategias bajo condiciones cambiantes. Dentro de este período se identifican cuatro subperíodos de estrés que estructuran el análisis comparativo.

El primero corresponde a las elecciones primarias PASO de 2019, analizado entre el 1 de junio y el 31 de octubre de ese año. El resultado electoral sorpresivo del 11 de agosto

desencadenó una de las caídas más abruptas de la historia reciente del mercado argentino: el índice Merval colapsó más de un 30% en una única jornada, el tipo de cambio oficial se disparó y el riesgo país superó los 2.000 puntos básicos en pocas horas, convirtiendo este episodio en un caso de prueba exigente para cualquier estrategia de gestión.

El segundo corresponde a la crisis del COVID-19, analizado entre el 1 de febrero y el 31 de julio de 2020. A diferencia del shock electoral de 2019, la pandemia representó un episodio de desarrollo más gradual: comenzó con señales de alerta desde enero, se intensificó con las primeras medidas de cuarentena global en marzo y se extendió durante varios meses con distintas oleadas de incertidumbre, permitiendo evaluar la capacidad de las estrategias para adaptarse a un régimen de alta volatilidad sostenida.

El tercero corresponde al período de crisis y elecciones de 2023, analizado entre el 1 de mayo y el 31 de diciembre de ese año. Este subperíodo se caracterizó por una combinación de aceleración inflacionaria, tensiones cambiarias extremas y un proceso electoral que derivó en un cambio de régimen económico y político. A diferencia de los episodios anteriores, incluyó también un fuerte rally de activos financieros locales impulsado por expectativas de cambio, generando una dinámica mixta de estrés y oportunidad simultánea especialmente útil para evaluar la capacidad de los modelos de capturar señales direccionales en un entorno de alta incertidumbre.

El cuarto corresponde al período electoral de 2025, analizado entre marzo y octubre de ese año. En un contexto de reconfiguración de expectativas económicas y financieras, este episodio representa un escenario de transición de mercado que permite evaluar si los modelos mantienen capacidad predictiva bajo condiciones distintas a las de los episodios anteriores, con historia de entrenamiento acumulada.

La selección de estos cuatro episodios cubre un espectro amplio de tipos de estrés: un shock político abrupto, una crisis sanitaria global de desarrollo gradual, una crisis macroeconómica local con componente electoral y un período de transición de régimen político-económico. Esta diversidad fortalece el diseño porque evita que las conclusiones dependan de las características de un único tipo de evento.

6.3 Universo de activos y fuentes de datos

El panel de activos sobre el cual se construyen las carteras algorítmicas está compuesto por 25 instrumentos de renta variable: 23 acciones que cotizan en Bolsas y Mercados

Argentinos (BYMA) y 2 American Depositary Receipts (ADRs), Vista Energy (VIST) y MercadoLibre (MELI), que cotizan en mercados estadounidenses. La inclusión de estos dos ADRs responde a la necesidad de incorporar activos con alta liquidez internacional, fuerte vinculación con la economía local y denominación en dólares, representativos del universo de inversión disponible para un fondo de renta variable argentina (ver Anexo 8). La selección priorizó activos con suficiente historia de precios durante todo el período de análisis y niveles de operatoria que garanticen su negociación en condiciones normales de mercado, excluyendo aquellos con operatoria esporádica o largos períodos sin cotización por el riesgo de introducir sesgos de liquidez *(Ver código en Anexo 8).*

Los precios diarios de cierre fueron obtenidos a través de la librería yfinance de Python, ajustados por dividendos y splits para garantizar la consistencia de las series temporales. Dado que los activos cotizan tanto en pesos como en dólares y que el período incluye episodios de fuerte devaluación, todos los retornos fueron expresados en dólares al tipo de cambio contado con liquidación (CCL), calculado a partir de datos de Bloomberg, homogeneización indispensable para evitar que las métricas de rendimiento queden distorsionadas por la inflación y la depreciación nominal del peso.

Como benchmark de la gestión humana activa se utilizaron los retornos de 12 fondos comunes de inversión de renta variable argentina: Superfondo Acciones A, Superfondo Renta Variable B, SBS Acciones ARG B, Galileo Acciones A, Rofex 20 Renta Variable B, CTIO Acciones ARG B, Schroder Renta Variable B, Adcap Acciones B, Balanz Acciones ARG B, Allaria Acciones B, FBA Acciones ARG B y Fima Acciones A. Los datos de cuotapartes fueron obtenidos de los reportes de 1816, consultora especializada en mercados financieros argentinos. Estos fondos representan los vehículos de mayor patrimonio administrado dentro del universo de renta variable local. Sus rendimientos son netos de comisiones y gastos de administración, lo que introduce una asimetría favorable a los modelos algorítmicos que se reconoce explícitamente como una limitación del análisis comparativo.

6.4 Variables e ingeniería de features

Los tres modelos de Machine Learning utilizados en esta investigación comparten el mismo conjunto de variables predictivas, construidas exclusivamente a partir de los precios históricos de cada activo. Esta decisión responde a dos criterios: mantener la comparabilidad entre modelos, dado que todos reciben exactamente la misma

información de entrada, y restringirse a variables disponibles en tiempo real al momento de generar cada predicción, garantizando la ausencia de lookahead bias.

Las variables utilizadas son cuatro. El retorno diario del activo constituye la variable base. El momentum a 5 días, calculado como el retorno acumulado en las últimas 5 jornadas, captura la tendencia de muy corto plazo. El momentum a 20 días, calculado de manera análoga, incorpora una perspectiva de mediano plazo que complementa la señal anterior. La volatilidad histórica a 20 días, calculada como la desviación estándar de los retornos diarios sobre esa ventana móvil, mide el nivel de incertidumbre reciente del activo *(ver código en Anexo 9).*

La elección de estas variables se sustenta en la evidencia empírica revisada en el marco teórico. Jegadeesh y Titman (1993) documentaron la persistencia del efecto momentum en horizontes de 3 a 12 meses, y Gu et al. (2020) identificaron variantes de momentum y volatilidad como las señales predictivas dominantes entre más de 900 variables candidatas. La selección deliberadamente reducida de features responde a la lógica de parsimonia: en mercados con datos limitados como el argentino, los modelos con menor número de variables tienden a generalizar mejor que aquellos sobreestimados con cientos de predictores de dudosa relevancia local.

6.5 Diseño de los modelos de Machine Learning

Los tres modelos implementados Random Forest, XGBoost y LSTM abordan el problema de inversión como una tarea de clasificación binaria: para cada activo del panel y para cada jornada de la simulación, el modelo estima la probabilidad de que el retorno del día siguiente sea positivo, probabilidad que actúa como señal de entrada para el optimizador de Markowitz (1952), que determina los pesos de la cartera sobre la base de las predicciones generadas.

El punto de corte entre entrenamiento y simulación se fijó en el 31 de mayo de 2019. Los modelos se entrenan con todos los datos históricos disponibles hasta esa fecha y no se reentrenan en ningún momento posterior, decisión deliberada que responde al objetivo de simular fielmente las condiciones de un gestor algorítmico real. La garantía de ausencia de lookahead bias, uno de los sesgos metodológicos más graves en investigaciones de este tipo según López de Prado (2018), se implementa de manera estricta: en cada jornada t los modelos generan predicciones utilizando únicamente información disponible hasta t-1, y las órdenes se ejecutan al precio de apertura del día t.

Random Forest se implementa con 500 árboles de decisión y una profundidad máxima de 10 niveles, evaluando en cada nodo un subconjunto aleatorio de variables *(ver código en Anexo 11)*. XGBoost se configura como clasificador binario con una tasa de aprendizaje de 0,05 y 200 iteraciones de boosting, incorporando regularización L1 y L2 para limitar la complejidad del modelo *(ver código en Anexo 12)*. Las redes LSTM se implementan con una capa recurrente de 64 unidades, una ventana temporal de 20 días y un dropout de 0,2 para evitar que el modelo memorice el histórico en lugar de aprender patrones generalizables *(ver código en Anexo 13)*. Los tres modelos fueron corridos 5 veces con distintas semillas aleatorias, reportándose el promedio de esas cinco corridas para verificar la consistencia de los resultados.

La implementación se llevó a cabo en Python, utilizando yfinance para la descarga de precios, pandas y numpy para el procesamiento de series temporales, scikit-learn para Random Forest, xgboost para el modelo de gradient boosting, y tensorflow con keras para las redes LSTM. La optimización de cartera bajo el criterio de Markowitz (1952) fue resuelta mediante scipy, incorporando las restricciones de peso máximo por activo y la condición de no permitir posiciones cortas.

6.6 Optimización de cartera (Markowitz)

Las probabilidades de retorno positivo generadas por cada modelo no se utilizan directamente como señal de trading, sino como insumo para un proceso de optimización de cartera bajo el marco de media-varianza propuesto por Markowitz (1952). El objetivo de la optimización es encontrar, para cada jornada, el vector de pesos que maximice el ratio de Sharpe esperado de la cartera, definido como el exceso de retorno esperado por unidad de riesgo asumido.

Las restricciones impuestas al optimizador son dos. Primero, no se permiten posiciones cortas: todos los pesos deben ser no negativos, lo que limita el universo de estrategias a carteras long-only consistentes con las restricciones operativas de los fondos comunes de inversión utilizados como benchmark. Segundo, ningún activo puede recibir un peso superior al 20% del total de la cartera, restricción que impone un mínimo de diversificación y evita que la solución se concentre en un único activo con la mayor probabilidad predicha. *(Ver código en anexo 10)*

La matriz de covarianzas utilizada en la optimización se estima sobre una ventana móvil de 60 días previos a cada jornada de rebalanceo. En línea con lo señalado por López de Prado (2018), en mercados con cambios de régimen frecuentes el uso de ventanas

cortas permite que la matriz refleje las condiciones de riesgo recientes con mayor precisión que estimaciones de largo plazo, que pueden incorporar patrones estructurales que ya no son relevantes al momento de la decisión.

6.7 Simulación del fondo

La simulación parte de un capital inicial de $1.000.000 ARS al 3 de junio de 2019 y opera con rebalanceo diario durante todas las jornadas hábiles hasta el 27 de febrero de 2026, abarcando un total de 1.642 jornadas de operatoria. En cada jornada el proceso sigue la siguiente lógica: los modelos generan las probabilidades de retorno positivo para cada activo usando información disponible hasta t-1, el optimizador de Markowitz determina los pesos óptimos de la cartera, y el capital se redistribuye según esos pesos al precio de apertura del día t. *(Ver código en anexo 14)*

La simulación no incorpora costos de transacción ni spread bid-ask. Esta decisión es una limitación explícita del trabajo y no una omisión involuntaria. En investigaciones de backtesting con rebalanceo diario, los costos de transacción pueden tener un impacto significativo sobre los resultados netos, especialmente en mercados con baja liquidez como el argentino. Sin embargo, su incorporación requeriría supuestos sobre el impacto de mercado de cada operación que, dada la ausencia de datos confiables para el período analizado, introducirían un grado de arbitrariedad metodológica mayor al que implica su exclusión.

6.8 Construcción de los promedios y criterio de comparación

La comparación central de esta investigación se realiza entre dos conjuntos agregados: el promedio de los tres modelos de Machine Learning por un lado, y el promedio de los doce fondos comunes de inversión por el otro. Esta decisión responde al objetivo del trabajo, que no es determinar cuál modelo individual es superior sino evaluar si la gestión algorítmica, en su conjunto, supera a la gestión humana activa

El promedio de los modelos de Machine Learning se construye calculando, para cada jornada, el promedio del valor de cartera de los tres modelos Random Forest, XGBoost y LSTM, donde cada modelo ya representa a su vez el promedio de sus cinco corridas con distintas semillas. De esta manera, el promedio ML refleja el desempeño esperado de un inversor que hubiera asignado capital de manera equitativa entre los tres enfoques algorítmicos.

El promedio de los fondos comunes de inversión se construye de manera análoga: para cada jornada se calcula el promedio del valor de cartera de los doce fondos del benchmark, partiendo todos de un capital inicial equivalente de $1.000.000 ARS. Esto permite que la comparación sea simétrica, ya que ambos conjuntos arrancan desde el mismo punto y se evalúan sobre la misma escala.

La comparación entre ambos promedios a lo largo del tiempo es la que permite responder la pregunta de investigación central: si el promedio ML supera consistentemente al promedio FCI durante el período completo y durante los episodios de estrés identificados, la evidencia apoya la hipótesis del trabajo.

6.9 Análisis de robustez

Para validar que los resultados no dependen de configuraciones aleatorias específicas, se realizaron 5 corridas independientes con semillas distintas para los tres modelos. En cada corrida se registran las métricas de evaluación y los resultados finales reportados corresponden al promedio de las 5 corridas junto con su rango de variación.

La consistencia de los resultados entre corridas es evidencia de que el desempeño observado no es un artefacto de una inicialización afortunada sino una propiedad estable de los modelos. Esta práctica responde a las recomendaciones de López de Prado (2018) sobre la validación de estrategias algorítmicas y permite descartar que los resultados sean producto del overfitting a una configuración particular del generador de números aleatorios.

6.10 Validación estadística de los resultados

El análisis estadístico de los resultados se desarrolla en dos etapas complementarias. En primer lugar, se aplica un análisis descriptivo sobre las series diarias de retornos del promedio ML y del promedio FCI, reportando media, mediana, desvío estándar, valores extremos, asimetría, curtosis, drawdown máximo, retorno acumulado y correlación entre ambas series, calculadas tanto para el período completo como para cada subperíodo de estrés. Este análisis permite caracterizar las propiedades distribucionales de las variables e identificar características relevantes para la interpretación de los tests inferenciales.

En segundo lugar, se aplicó un paquete de cuatro tests complementarios sobre las series diarias de retornos para evaluar formalmente si las diferencias observadas son estadísticamente significativas. La elección de este conjunto responde a la lógica de

contrastar cuatro preguntas distintas que, en conjunto, validan la solidez de los hallazgos: si los modelos generan retornos positivos, si superan al benchmark, si esa superioridad se mantiene ajustando por riesgo, y si los resultados son robustos al remuestreo.

El primer test evalúa si los retornos diarios de cada estrategia son estadísticamente distintos de cero mediante un test t de una muestra, contrastando H_0: μ = 0 contra H_1: μ > 0 en versión unilateral. Su aplicación a ambas series por separado permite identificar cuál de las dos estrategias genera retornos sistemáticamente positivos en términos formales.

El segundo test, núcleo del paquete inferencial, contrasta si los modelos ML presentan desempeño superior al de los FCI mediante un test t pareado sobre la diferencia diaria (ret_ML − ret_FCI), bajo H_0: μ_diff = 0 contra H_1: μ_diff > 0. La elección del test pareado responde a que ambas series comparten las mismas jornadas y los mismos shocks de mercado, por lo que un test independiente subestimaría la potencia estadística al ignorar la correlación entre series. Como prueba de robustez no paramétrica se aplica adicionalmente el test de rangos con signo de Wilcoxon (1945) sobre la misma diferencia diaria, que no requiere el supuesto de normalidad frecuentemente violado en series financieras con colas pesadas. La concordancia entre ambos tests refuerza la solidez de las conclusiones; una divergencia indicaría que el resultado paramétrico depende de pocas observaciones extremas.

El tercer test valida si la superioridad del promedio ML se mantiene una vez ajustada por volatilidad, aplicando el test de Memmel (2003), corrección del test original de Jobson y Korkie (1981), diseñado para comparar ratios de Sharpe sobre series correlacionadas. El test contrasta H_0: Sharpe_ML = Sharpe_FCI contra H_1: Sharpe_ML > Sharpe_FCI mediante un estadístico z asintóticamente normal que incorpora la correlación entre series en el cálculo de la varianza, apropiado dado que ambas estrategias se aplican al mismo mercado. Su inclusión es especialmente relevante dado que la optimización de Markowitz empleada en este trabajo utiliza el ratio de Sharpe como criterio de selección de cartera.

Como cuarta capa de validación se construyen intervalos de confianza al 95% sobre el retorno acumulado de cada estrategia mediante bootstrap no paramétrico, siguiendo el procedimiento de Efron y Tibshirani (1993), con diez mil remuestreos con reemplazo. A

diferencia de los tests paramétricos, el bootstrap no asume ninguna distribución particular ni depende de aproximaciones asintóticas, complementando los tests anteriores con una medida de robustez basada exclusivamente en la información empírica observada.

Los cuatro tests se aplican tanto al período completo como a cada uno de los cuatro subperíodos de estrés. Cabe señalar que la menor cantidad de observaciones en cada subperíodo entre ciento dos y ciento sesenta y tres jornadas reduce la potencia estadística, por lo que los resultados por subperíodo deben interpretarse considerando esta limitación inherente al tamaño de la muestra.

7. Resultados

7.1 Período completo

El análisis del período completo, comprendido entre el 3 de junio de 2019 y el 27 de febrero de 2026, revela una superioridad sostenida del promedio ML frente al promedio FCI a lo largo de más de seis años y medio de operatoria continua. Partiendo de un capital inicial equivalente a USD 22.166,50, el promedio ML alcanzó al cierre del período un retorno acumulado del 632%, frente al 117% del promedio FCI, una diferencia de 515 puntos porcentuales que constituye el principal hallazgo cuantitativo del trabajo. *(Ver grafico en Anexo 1)*

La evolución del comparativo no fue uniforme: durante los primeros años ambas estrategias se movieron en niveles similares en un mercado estancado, pero a partir de 2023, con el rally electoral y el cambio de régimen económico, los modelos ML capturaron de manera sistemática una mayor proporción de los movimientos alcistas, ampliando progresivamente la brecha hasta llevarla a más de cinco veces el retorno acumulado del benchmark humano. Esta superioridad se construyó sin asumir mayor riesgo: la volatilidad diaria de ambas estrategias resultó prácticamente idéntica y el drawdown máximo del promedio ML fue diez puntos porcentuales menor. El detalle de cómo se construyó esta ventaja durante cada uno de los cuatro episodios de estrés analizados se desarrolla en las secciones siguientes.

7.2 PASO 2019

El episodio de las elecciones primarias de agosto de 2019 representa el caso más extremo de shock abrupto del período analizado. En las semanas previas al 11 de

agosto, ambos enfoques venían acumulando retornos positivos similares, moviéndose prácticamente en sintonía en torno al +23% desde el inicio del período (ver Anexo 2).

El día 12 de agosto, primera jornada de operatoria tras el resultado electoral, el promedio ML cayó un 41,0% y el promedio FCI un 44,1%. La diferencia de apenas 3 puntos porcentuales en el día del shock indica que ninguno de los dos enfoques logró anticipar ni evitar el golpe inicial: ambos cayeron con fuerza porque el evento fue genuinamente imprevisible a partir de datos históricos de precios.

La diferencia de 12,8 puntos porcentuales al cierre del episodio ML en -28,3% contra FCI en -41,1% no se construyó el día del shock sino en las semanas siguientes. A partir del 13 de agosto el promedio ML comenzó a recuperarse con mayor velocidad: el 15 de agosto rebotó un 9,2% contra un 6,0% de los FCI, el 16 de agosto subió un 12,8% contra un 7,7%, y esta dinámica se repitió en los rebounds sucesivos. La ventaja algorítmica en este episodio no fue de anticipación sino de rebalanceo: los modelos lograron rotar hacia los activos con mayor señal de recuperación con mayor velocidad y disciplina que los gestores humanos.

7.3 COVID 2020

La crisis del COVID-19 presentó una dinámica completamente distinta al episodio anterior. En lugar de un shock concentrado en un día, la caída se desarrolló en varias semanas a medida que la pandemia se extendía globalmente. Los días más críticos fueron el 9 de marzo (-17,2% ML / -17,8% FCI), el 12 de marzo (-11,0% / -10,7%), el 16 de marzo (-11,1% / -10,9%) y el 18 de marzo (-18,9% / -19,3%), acumulando una caída del -51,5% para ML y -52,1% para FCI en ese punto más bajo. *(Ver grafico en Anexo 3)*

Al igual que en las PASO, ambos enfoques cayeron de manera muy similar durante los días de mayor pánico. La diferencia de rendimiento entre los dos tipos de gestión, ML en -11,1% contra FCI en -22,7% al cierre del episodio, se explica nuevamente por lo que ocurrió después del piso: desde fines de marzo en adelante, los modelos ML recuperaron terreno más rápido y de manera más sostenida que los fondos humanos. Este patrón sugiere que la ventaja algorítmica no radica en la capacidad de predecir shocks sino en la de posicionarse mejor durante la recuperación, rotando sistemáticamente hacia los activos con mayor momentum positivo sin los sesgos conductuales, como la aversión a las pérdidas o el anclaje al precio de entrada, que pueden llevar a un gestor humano a mantener posiciones perdedoras por más tiempo del conveniente.

7.4 Elecciones 2023

El episodio de 2023 es el más complejo de los cuatro analizados: no se trata de una caída seguida de recuperación sino de un período con múltiples cambios de liderazgo, semanas de fuerte volatilidad y un desenlace donde los modelos ML terminaron imponiendo una ventaja significativa que no era evidente durante gran parte del recorrido.

El episodio arranca en mayo con ambos enfoques moviéndose de manera muy parecida. Durante junio y julio los modelos ML construyeron una ventaja de entre 3 y 5 puntos porcentuales que parecía consolidarse. Sin embargo, entre principios de agosto y fines de octubre los FCI pasaron al frente: el mercado vivió un rally impulsado por las expectativas electorales que los fondos humanos capturaron mejor, llegando a superar a los modelos ML por hasta 16 puntos porcentuales hacia mediados de octubre el único período extendido del episodio en que los gestores humanos llevaron ventaja (ver Anexo 4).

El momento más crítico fue el 23 de octubre, primera jornada tras las elecciones generales que dejaron definida la disputa entre Massa y Milei para el ballotage. La incertidumbre generó una venta masiva: los FCI cayeron un 8,0% mientras que los modelos ML cayeron apenas un 0,6%, una diferencia de 7,4 puntos en un único día que recortó la brecha de -15,5 puntos a -6,5 puntos. La diferencia no se explica por anticipación sino por la lógica sistemática de los modelos: sin un gestor humano que reaccionara emocionalmente vendiendo, los algoritmos mantuvieron la composición de cartera que las señales de precio indicaban, absorbiendo el shock con mucha menor exposición.

El punto de inflexión definitivo llegó el 24 de noviembre, cuando Milei ganó el ballotage: los modelos ML subieron un 20,2% contra un 14,3% de los FCI, recuperando la totalidad de la desventaja acumulada y pasando al frente por primera vez desde agosto. El 27 de noviembre la brecha se consolidó con los ML sumando otro 2,4% mientras los FCI caían 2,6%, sellando el cambio de liderazgo con una diferencia de 4,5 puntos. Desde ese momento los modelos ML no volvieron a ceder el liderazgo, cerrando el episodio con un retorno acumulado del 63,6% para ML contra el 44,3% de los FCI, una brecha final de 19,3 puntos porcentuales.

7.5 Elecciones 2025

El episodio comenzó en abril con un shock externo: el anuncio de aranceles por parte de la administración Trump generó una corrección abrupta en los mercados globales que impactó de lleno en el mercado local. El 4 de abril ML cayó 8,7% y FCI 9,4% en una sola jornada, con ambos enfoques moviéndose en sintonía ante un evento de origen externo que no distinguió entre estrategias de gestión. *(Ver grafico en Anexo 5)*

El primer evento electoral fue las elecciones provinciales bonaerenses del 7 de septiembre. En las semanas previas el mercado ya venía incorporando en los precios una expectativa de resultado desfavorable para el oficialismo, lo que explica la tendencia bajista sostenida desde mayo. Lejos de revertirse tras la elección, esa tendencia se profundizó: el 8 de septiembre, primera jornada posterior al resultado, el mercado registró su peor caída del episodio con ML en -14,3% y FCI en -17,2%. La señal política que el resultado enviaba sobre el escenario nacional mantuvo al mercado en terreno negativo durante las semanas siguientes, llevando a ambos enfoques a pérdidas acumuladas superiores al 40%. En toda esa fase bajista los modelos ML sostuvieron una ventaja de entre 4 y 6 puntos sobre los FCI, absorbiendo mejor las caídas diarias de manera consistente.

El punto de inflexión llegó con las elecciones legislativas nacionales del 26 de octubre. El resultado favorable al oficialismo despejó la incertidumbre que había pesado sobre el mercado durante meses y desencadenó un rally explosivo: el 27 de octubre ML subió 34,4% y FCI 29,3% en una única jornada, con los algoritmos capturando nuevamente una mayor proporción del movimiento alcista, al igual que lo ocurrido tras el ballotage de 2023. El episodio cerró con ML en +12,3% y FCI en +5,7%.

7.6 Análisis descriptivo de las series

Antes de abordar el análisis inferencial, resulta necesario caracterizar estadísticamente las dos series diarias de retornos del promedio ML y del promedio FCI. La estadística descriptiva permite dimensionar las propiedades centrales, de dispersión y de forma de cada serie, lo que constituye una base imprescindible para evaluar luego la significancia estadística de las diferencias observadas y para interpretar correctamente los resultados financieros del comparativo.

Comenzando por la caracterización del período completo, la siguiente tabla sintetiza las principales medidas descriptivas calculadas sobre la serie de retornos diarios para el

lapso comprendido entre el 4 de junio de 2019 y el 27 de febrero de 2026, abarcando un total de 1.634 jornadas de operatoria.

(Ver tabla en Anexo 6)

La lectura de estos descriptivos permite identificar cuatro hechos que estructuran el resto del análisis.

En primer lugar, la volatilidad de ambas series es prácticamente idéntica: 3,41% diario para el promedio ML contra 3,45% del promedio FCI, equivalentes a aproximadamente 54% anual en ambos casos. Esta paridad en el desvío estándar implica que la mayor rentabilidad del promedio ML no proviene de asumir mayor riesgo absoluto, sino de una mejor combinación dentro del mismo nivel de riesgo. En otras palabras, ambos enfoques de gestión operan con la misma exposición a la volatilidad del mercado argentino pero obtienen retornos sistemáticamente diferentes, lo cual constituye en sí mismo evidencia preliminar de una asimetría real en la capacidad de capturar señales del mercado.

En segundo lugar, la correlación entre ambas series alcanza 0,957, valor extremadamente alto que confirma que ambas estrategias responden esencialmente a los mismos shocks de mercado. Esta característica se traduce en una elevada sincronía durante las jornadas de pánico y de euforia, donde ambas series se mueven en el mismo sentido aunque con magnitudes diferentes, lo que evidencia que la fuente principal de variabilidad es compartida y proviene de la dinámica agregada del mercado argentino.

En tercer lugar, las distribuciones de retornos se alejan significativamente de la normalidad, lo cual era esperable en series financieras de mercados emergentes. La asimetría negativa, particularmente pronunciada en el promedio FCI (-0,83 contra -0,19 del ML), indica una mayor concentración de retornos extremos a la baja. La curtosis exceso de aproximadamente 24 en ambas series, frente a un valor de cero en distribuciones normales, evidencia colas extremadamente pesadas: eventos extremos como las PASO 2019 o el COVID-19 ocurren con frecuencia muchísimo mayor a la que predeciría un modelo gaussiano.

En cuarto lugar, el drawdown máximo del promedio ML resulta diez puntos porcentuales menor al del promedio FCI (66,2% contra 76,3%). Si bien ambas estrategias sufrieron pérdidas catastróficas durante los peores momentos del período analizado, los modelos algorítmicos lograron limitar mejor la magnitud del derrumbe acumulado, evidencia

consistente con la hipótesis de que su lógica sistemática actúa como protección frente a las caídas extremas durante escenarios de pánico de mercado.

Aplicando el mismo análisis a los subperíodos de estrés, la siguiente tabla presenta una síntesis comparativa de los principales descriptivos durante los cuatro episodios identificados.

(Ver tabla en Anexo 6.1)

Tres patrones generales emergen del análisis por subperíodos.

Primero, el promedio ML supera al promedio FCI en retorno acumulado durante los cuatro subperíodos analizados, con ventajas que van desde 6,6 puntos porcentuales en Elecciones 2025 hasta 19,3 puntos en Elecciones 2023. El signo es consistente y los órdenes de magnitud son económicamente significativos en todos los casos.

Segundo, la volatilidad de ambas estrategias es nuevamente muy similar dentro de cada subperíodo, con diferencias entre ML y FCI menores a 30 puntos básicos en todos los casos. Este patrón refuerza la conclusión del análisis del período completo: la mejor rentabilidad del promedio ML no se explica por mayor exposición al riesgo, sino por una capacidad sistemática de capturar mejor el componente direccional del mercado.

Tercero, los valores extremos diarios evolucionan en sintonía entre ambas estrategias en los cuatro subperíodos, mientras que el drawdown máximo muestra un patrón mixto: el promedio ML registra un drawdown menor o similar al FCI en los episodios de PASO 2019, COVID 2020 y Elecciones 2025, pero un drawdown algo mayor en Elecciones 2023 (32,8% contra 25,3%). Este último caso es consistente con la dinámica del rally de mediados de 2023, donde los FCI lideraron mientras los modelos ML construían posiciones que se monetizarían recién tras el ballotage. La diferencia en retorno acumulado se construye, entonces, durante las jornadas posteriores a los shocks, momentos en los que los modelos algorítmicos capturan más rápido y de manera más sistemática las recuperaciones del mercado.

7.7 Validación estadística de los resultados

Para evaluar formalmente si las diferencias observadas entre el promedio ML y el promedio FCI son estadísticamente significativas o si podrían explicarse por azar, se aplicó un paquete de tests inferenciales cuyos resultados se reportan a continuación. La estructura del análisis sigue una lógica progresiva que comienza por verificar si cada estrategia genera retornos no aleatorios, avanza hacia la comparación directa entre

ambas, evalúa luego si la superioridad observada se mantiene ajustando por riesgo, y culmina con un análisis de robustez basado en remuestreo.

El primer paso consiste en evaluar si los retornos diarios de cada estrategia son estadísticamente distintos de cero mediante un test t de una muestra unilateral, donde la hipótesis alternativa plantea que el retorno medio es positivo (H_1: $\mu > 0$). El promedio ML genera un retorno diario medio del 0,18% con significancia estadística al 5% ($p = 0,017$), confirmando que la rentabilidad observada constituye evidencia de capacidad predictiva real y no de fluctuaciones aleatorias en torno a cero. El promedio FCI, por su parte, presenta un retorno diario del 0,11% con un p-valor de 0,106, que no permite rechazar formalmente la hipótesis nula al 5%, aunque el signo de la media es positivo.

Confirmado que el promedio ML genera retornos significativamente positivos, el siguiente paso consiste en evaluar si efectivamente supera al promedio FCI, contraste que constituye el núcleo de la validación del trabajo. En el período completo, el test t pareado rechaza con claridad la hipótesis de igualdad de medias ($t = 2,92$, $p = 0,0018$), confirmando que el promedio ML supera al promedio FCI con alta significancia estadística. La diferencia diaria promedio entre ambas estrategias es de 7,2 puntos básicos, equivalente a aproximadamente 18 puntos porcentuales de ventaja anual a favor del enfoque algorítmico. El test de Wilcoxon, más conservador, arroja un p-valor de 0,059, levemente por encima del umbral del 5%. La leve divergencia entre ambos tests refleja una característica de la dinámica del comparativo: la ventaja del enfoque algorítmico se concentra en los rebotes posteriores a shocks de mercado, jornadas extremas que el test paramétrico captura plenamente y que el test basado en rangos modera al asignarles menor peso relativo.

En los cuatro subperíodos de estrés analizados individualmente, el signo de la diferencia es favorable al promedio ML en todos los casos, con ventajas que van desde 3,5 puntos básicos diarios en Elecciones 2025 hasta 17,3 en PASO 2019. Sin embargo, los p-valores no alcanzan significancia estadística al 5% en ningún subperíodo. Este resultado responde a una limitación esperable: con muestras de entre 102 y 163 jornadas por subperíodo, la potencia estadística disponible es insuficiente para detectar diferencias de medias en horizontes tan cortos. La conclusión a extraer no es que la ventaja del enfoque algorítmico no exista en estos episodios, sino que el horizonte necesario para validarla formalmente excede la duración de cada uno por separado.

La diferencia en retornos medios podría reflejar, sin embargo, una mayor exposición al riesgo por parte del enfoque algorítmico. Para descartar esa hipótesis alternativa se aplicó el test de Memmel sobre la diferencia de ratios de Sharpe, que confirma una superioridad sólida en términos ajustados por riesgo. El ratio de Sharpe anualizado del promedio ML alcanza 0,83 frente a 0,49 del promedio FCI, una diferencia altamente significativa ($z = 2,97$, $p = 0,0015$) que constituye, junto con el test t pareado, el resultado estadístico más robusto del trabajo. La interpretación es directa: los modelos algorítmicos no solo generan mayor rentabilidad sino que la generan con una relación retorno-riesgo significativamente más eficiente que el promedio de los FCI. En los subperíodos individuales el signo nuevamente es favorable al promedio ML en todos los casos, pero los p-valores tampoco alcanzan significancia formal por la misma limitación de tamaño muestral.

Como capa adicional de validación se construyeron intervalos de confianza al 95% sobre el retorno acumulado de cada estrategia mediante bootstrap no paramétrico. Los intervalos resultantes son amplios debido a la alta volatilidad de las series financieras y al efecto multiplicativo del retorno compuesto sobre horizontes de seis años, fenómeno esperable y bien documentado en la literatura. Por esa razón, los tests paramétricos basados en la distribución de la media (tests t y de Memmel) resultan más informativos para evaluar la significancia de la diferencia entre las dos estrategias y son los que sostienen las conclusiones formales del trabajo. El bootstrap se incluye como una capa complementaria de robustez antes que como evidencia decisoria.

La siguiente tabla sintetiza los resultados de los cinco contrastes aplicados sobre el período completo.

(Ver tabla en Anexo 7)

En conjunto, la evidencia estadística sustenta la hipótesis central del trabajo en el período completo 2019-2026: los modelos de machine learning superaron al promedio de los FCI tanto en retorno medio diario como en ratio de Sharpe, con significancia estadística al 1% en ambos casos. En los subperíodos de estrés individuales la ventaja se mantiene en signo en todos los casos pero no alcanza significancia formal, limitación esperable dada la baja potencia estadística asociada a muestras de 102 a 163 observaciones. El hallazgo central del trabajo, el desempeño superior de la gestión algorítmica frente a la gestión humana en el mercado argentino, queda así respaldado por evidencia estadística formal.

8. Conclusiones

El presente trabajo se propuso analizar el desempeño comparativo entre estrategias de gestión algorítmica y gestión humana activa en el mercado argentino durante el período 2019–2026, con foco en episodios de estrés financiero. A lo largo del trabajo se construyeron y evaluaron tres modelos de Machine Learning, Random Forest, XGBoost y LSTM, combinados con optimización de cartera bajo el criterio de Markowitz, comparando su desempeño contra doce fondos comunes de inversión de renta variable argentina como representantes de la gestión humana activa.

La evidencia empírica obtenida confirma la hipótesis planteada en la presente investigación: se observan diferencias significativas y consistentes a favor de las estrategias basadas en machine learning respecto de la gestión activa humana, particularmente en episodios de estrés financiero. En los cuatro episodios de estrés analizados, PASO 2019, COVID 2020, Elecciones 2023 y Elecciones 2025, el promedio ML superó al promedio de los FCI, y esta superioridad se sostuvo también en el período completo. Partiendo de un capital equivalente a USD 22.166,50, el promedio ML alcanzó USD 162.207,87 al cierre, equivalente a un retorno acumulado del 632% en dólares, frente al 117% del promedio de los doce FCI del benchmark. Un hallazgo transversal a los cuatro episodios merece especial atención: la ventaja algorítmica no proviene de anticipar los shocks sino de reaccionar mejor después de ellos. En el PASO 2019 ambos enfoques cayeron con fuerza el día del evento, pero los modelos ML se recuperaron más rápido en las semanas siguientes. En el COVID el piso fue similar para ambos pero la recuperación algorítmica fue más veloz y sostenida. En las Elecciones 2023 los FCI incluso llevaron ventaja durante varios meses, pero los modelos revertieron esa desventaja en jornadas puntuales de alta volatilidad donde su lógica sistemática los protegió mejor. En las Elecciones 2025 la ventaja fue más gradual pero consistente a lo largo de toda la fase bajista.

Este resultado no es aislado sino que emerge de la consistencia interna entre las distintas partes del trabajo. El marco teórico identificó en los sesgos conductuales, aversión a las pérdidas, anclaje y exceso de confianza, las principales fuentes de deterioro de la gestión humana bajo incertidumbre, y la evidencia empírica lo confirma: los momentos donde la brecha se amplió más a favor de los algoritmos coinciden precisamente con las jornadas de mayor pánico o euforia, donde esos sesgos operan con mayor intensidad. A su vez, las ineficiencias estructurales del mercado argentino

documentadas por Bekaert y Harvey (2002) y el marco de Grossman y Stiglitz (1980) anticipaban que un sistema capaz de procesar información de manera sistemática y a bajo costo marginal tendría ventajas en este contexto, y los resultados son consistentes con esa predicción teórica.

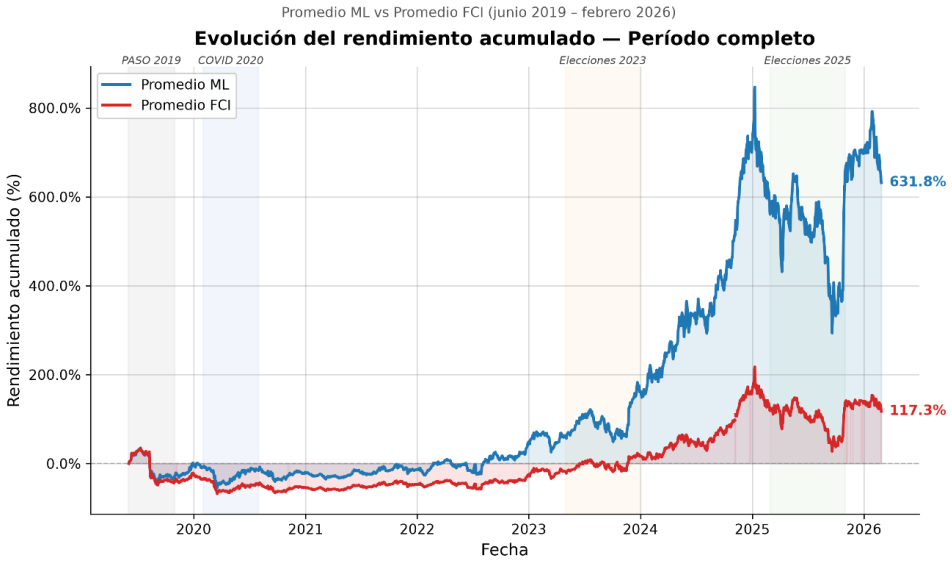
En términos de aporte al conocimiento, este trabajo extiende la evidencia de Harvey et al. (2017) y Grobys et al. (2022), que compararon gestión humana y algorítmica en mercados desarrollados, hacia un contexto que esos estudios dejaron inexplorado. Los resultados sugieren que la ventaja relativa de los modelos algorítmicos puede ser incluso mayor en mercados emergentes como el argentino, donde el margen de ineficiencias explotables es estructuralmente más amplio. Este hallazgo contribuye a cerrar el vacío identificado en la literatura y abre una línea de investigación con potencial de generalización a otros mercados emergentes con características similares.

Las principales limitaciones del trabajo deben reconocerse explícitamente. La ausencia de costos de transacción en la simulación es la más relevante: con rebalanceo diario en un mercado de liquidez acotada, los costos reales podrían reducir de manera significativa los retornos netos de las estrategias algorítmicas. A esto se suma la asimetría de comparación: los rendimientos de los FCI son netos de comisiones mientras que los modelos son brutos, lo que favorece mecánicamente a los algoritmos. Finalmente, los resultados corresponden a un backtesting histórico y no garantizan replicabilidad en períodos futuros. Las líneas de investigación futura más relevantes son incorporar costos de transacción realistas, ampliar el panel con variables macroeconómicas como features adicionales y extender el análisis a otros mercados emergentes.

Desde el punto de vista práctico, los resultados tienen implicaciones concretas para inversores institucionales, gestores de fondos y reguladores que operan en mercados con alta volatilidad estructural. La evidencia sugiere que la incorporación de herramientas algorítmicas en los procesos de inversión puede mejorar el desempeño relativo especialmente en contextos de estrés, donde los sesgos conductuales operan con mayor intensidad y la disciplina sistemática de los algoritmos representa una ventaja estructural difícil de replicar mediante la gestión discrecional.

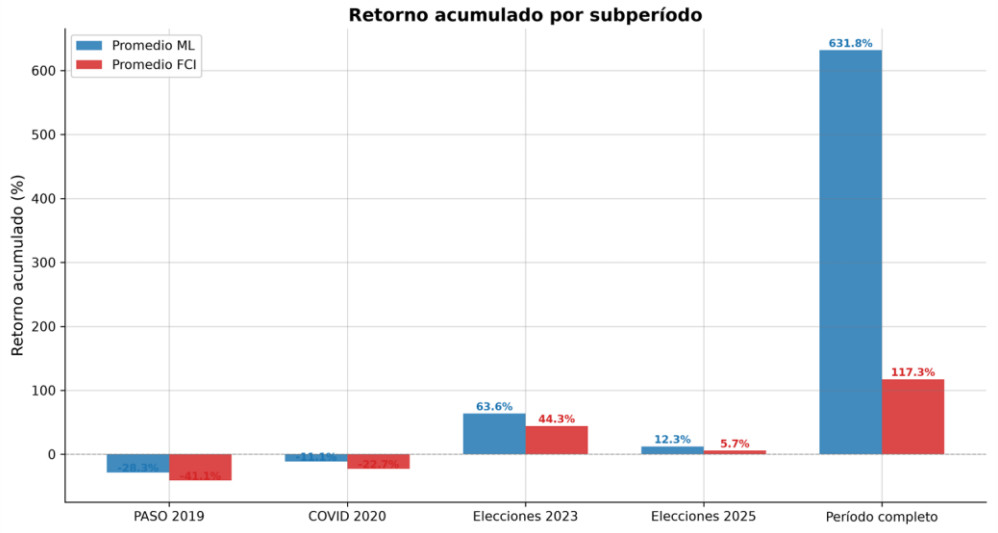
Anexos

Anexo 1



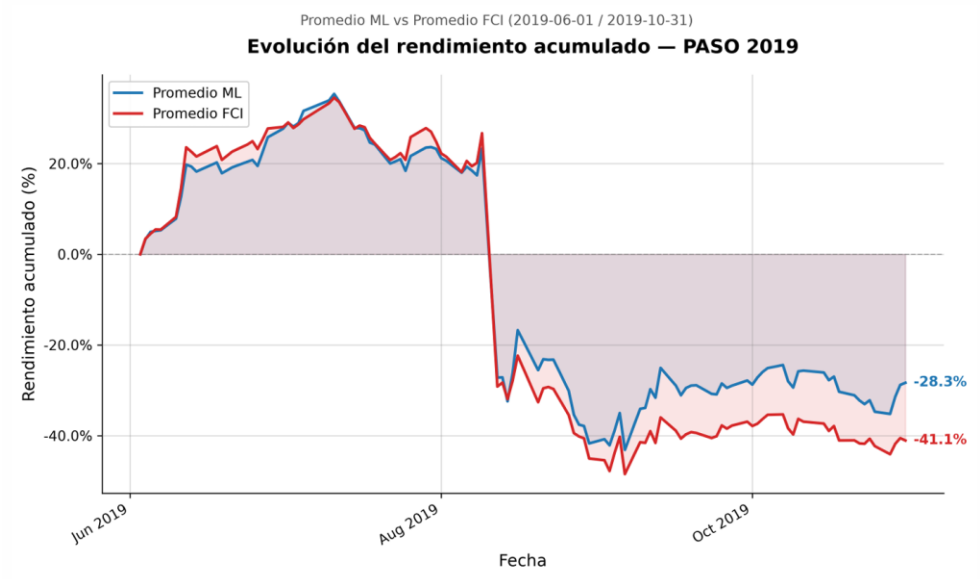
Evolución del valor de cartera del promedio ML y el promedio FCI durante el período junio 2019 – febrero 2026. Las bandas sombreadas indican los cuatro subperíodos de estrés financiero analizados. Fuente: elaboración propia.

Anexo 1.1



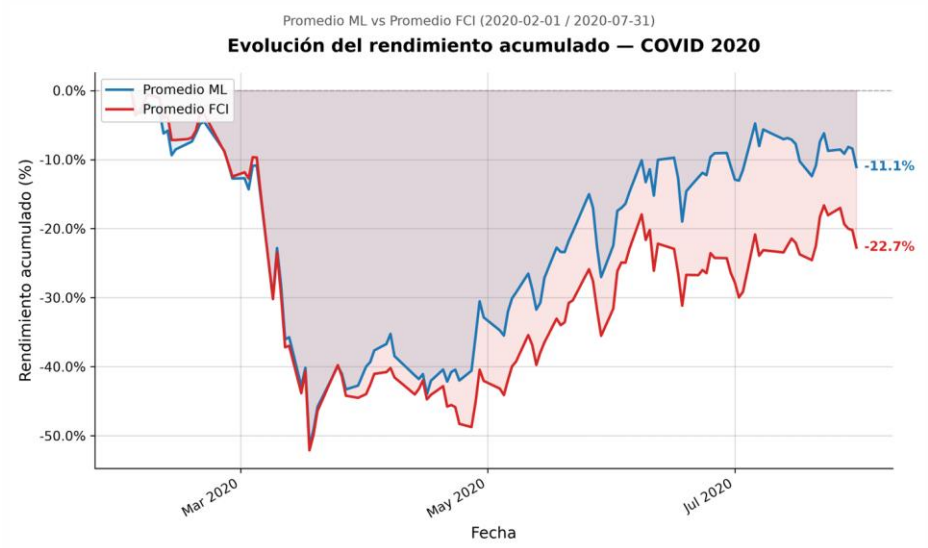
Retorno acumulado del promedio ML y el promedio FCI durante cada subperíodo de estrés analizado y durante el período completo. Las etiquetas sobre cada barra indican el valor numérico correspondiente. Fuente: elaboración propia.

Anexo 2



Evolución del rendimiento acumulado del promedio ML y el promedio FCI durante el subperíodo correspondiente a las elecciones primarias PASO 2019 (1 de junio – 31 de octubre de 2019). El sombreado refleja la diferencia acumulada entre ambas estrategias. Fuente: elaboración propia.

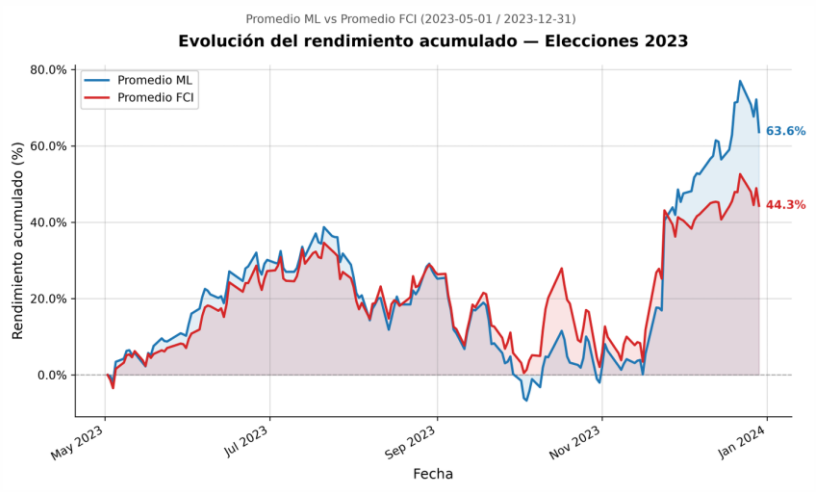
Anexo 3



Evolución del rendimiento acumulado del promedio ML y el promedio FCI durante el subperíodo correspondiente a la crisis del COVID-19 (1 de febrero – 31 de julio de 2020). El sombreado refleja la diferencia

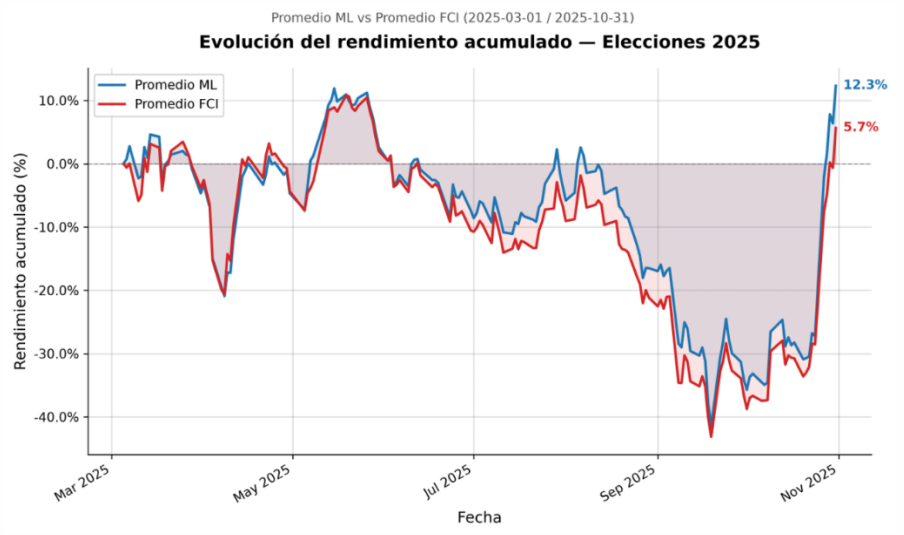
acumulada entre ambas estrategias. Fuente: elaboración propia.

Anexo 4



Evolución del rendimiento acumulado del promedio ML y el promedio FCI durante el subperíodo correspondiente a la crisis y elecciones de 2023 (1 de mayo – 31 de diciembre de 2023). El sombreado refleja la diferencia acumulada entre ambas estrategias. Fuente: elaboración propia.

Anexo 5



Evolución del rendimiento acumulado del promedio ML y el promedio FCI durante el subperíodo correspondiente a las elecciones de 2025 (1 de marzo – 31 de octubre de 2025). El sombreado refleja la diferencia acumulada entre ambas estrategias. Fuente: elaboración propia.

Anexo 6

Métrica	Promedio ML	Promedio FCI
Observaciones (n)	1.634	1.634
Media diaria	0,18%	0,10%
Mediana diaria	0,21%	0,12%
Desvío estándar diario	3,41%	3,45%
Desvío estándar anualizado	54,17%	54,73%
Mínimo diario	-40,97%	-44,10%
Máximo diario	34,39%	29,28%
Asimetría	-0,19	-0,83
Curtosis (exceso)	23,90	23,78
Drawdown máximo	-66,18%	-76,26%
Retorno acumulado	631,77%	117,34%
Correlación ML / FCI	0,957	

Anexo 6.1

Subperíodo	Estrategia	n	Media diaria	Desvío diario	Ret. acumulado	Max drawdown
PASO 2019	ML	102	-0,131%	5,76%	-30,63%	-58,00%
	FCI	102	-0,303%	5,89%	-43,02%	-61,70%
COVID 2020	ML	118	0,024%	4,56%	-11,08%	-51,49%
	FCI	118	-0,082%	4,77%	-22,74%	-52,11%
Elecciones 2023	ML	163	0,340%	3,20%	63,62%	-32,76%
	FCI	163	0,258%	3,16%	44,31%	-25,34%
Elecciones 2025	ML	162	0,177%	4,53%	12,32%	-47,67%

	FCI	162	0,142%	4,43%	5,67%	-48,71%

Síntesis comparativa de los principales descriptivos del promedio ML y el promedio FCI durante los cuatro subperíodos de estrés analizados. Se reportan media diaria, desvío diario, retorno acumulado y drawdown máximo de cada estrategia. Fuente: elaboración propia.

Anexo 7

Hipótesis evaluada	**Test aplicado**	**Resultado período completo**
Generación de retornos positivos del promedio ML	t de una muestra	Sí, p = 0,017 ∗
Generación de retornos positivos del promedio FCI	t de una muestra	No, p = 0,106
Superioridad del promedio ML en media diaria	t pareado	Sí, p = 0,0018 ∗∗
Superioridad del promedio ML (robustez no paramétrica)	Wilcoxon	Borderline, p = 0,059
Superioridad del promedio ML ajustado por riesgo	Memmel sobre Sharpe	Sí, p = 0,0015 ∗∗

(∗ p < 0,05; ∗∗ p < 0,01)

Resumen de los resultados de los cinco contrastes estadísticos aplicados sobre el período completo. Fuente: elaboración propia.

Anexo 8

Definición del panel de 25 activos

El presente fragmento corresponde al script de la Etapa 1 e indica el conjunto de activos sobre los cuales se construyó el panel, los identificadores utilizados para cada uno en la fuente Yahoo Finance, su clasificación sectorial y el rango temporal cubierto por la descarga. Esta selección constituye el punto de partida del trabajo y define el universo sobre el cual operan tanto los modelos algorítmicos como los fondos comunes de inversión utilizados como benchmark.

```python
# ============================================================
# PANEL: 23 acciones locales en ARS (BYMA) + 2 ADRs en USD
# Período: 2019-01-01 → 2026-03-01
# ============================================================

PANEL_LOCAL = {
    # Financiero
    'GGAL.BA':  'Grupo Galicia',
    'BMA.BA':   'Banco Macro',
    'BBAR.BA':  'BBVA Argentina',
    'SUPV.BA':  'Grupo Supervielle',
    'VALO.BA':  'Grupo Financiero Valores',
    # Energia
    'PAMP.BA':  'Pampa Energia',
    'CEPU.BA':  'Central Puerto',
    'TRAN.BA':  'Transener',
    'TGNO4.BA': 'Transportadora Gas del Norte',
    'EDN.BA':   'Edenor',
    'TGSU2.BA': 'TGS',
    'YPFD.BA':  'YPF',
    'METR.BA':  'Metrogas',
    # Industrial / Materiales
    'ALUA.BA':  'Aluar',
    'TXAR.BA':  'Ternium Argentina',
    'LOMA.BA':  'Loma Negra',
    'MIRG.BA':  'Mirgor',
    # Agro / Holding
    'CRES.BA':  'Cresud',
    'COME.BA':  'Comercial del Plata',
    'MORI.BA':  'Morixe Hermanos',
    # Telecom / Media
    'TECO2.BA': 'Telecom',
    'CVH.BA':   'Cablevision Holding',
    # Mercado de capitales
    'BYMA.BA':  'BYMA',
}

# ADRs en USD: se descargan en dolares, se guardan en USD
# La conversion a ARS via CCL se hace en Etapa 2 con datos Bloomberg
PANEL_ADR = {
    'VIST': ('Vista Energy',  '2020-01-01'),  # NYSE desde 2019, BYMA desde 2020
    'MELI': ('MercadoLibre',  '2019-01-01'),  # NASDAQ desde 1999
}

# Períodos de estrés
periodos = {
    'PASO_2019':        ('2019-06-01', '2019-10-31'),
    'COVID_2020':       ('2020-02-01', '2020-07-31'),
    'CRISIS_2023':      ('2023-05-01', '2023-12-31'),
    'ELECCIONES_2025':  ('2025-03-01', '2025-10-31'),  # PASO 27/04 + Generales 26/10
}
FECHA_INICIO = '2019-01-01'
FECHA_FIN    = '2026-03-01'
```

Fuente: elaboración propia.

Anexo 9

Cálculo de variables predictivas: momentum y volatilidad

El presente fragmento corresponde al script de la Etapa 1 y muestra la construcción de las variables predictivas utilizadas como insumo de los modelos de machine learning. Para cada activo del panel se calculan retornos diarios, volatilidad histórica móvil a veinte días, y momentum a cinco y veinte días, calculados como suma móvil de los retornos diarios. La construcción se realiza por separado para acciones locales en pesos y para los ADRs en dólares.

```
# ── Variables financieras — en ARS para locales, USD para ADRs ──
acciones_locales = list(precios_ars_raw.columns)

retornos_ars = precios_ars[acciones_locales].pct_change()
vol_20d      = retornos_ars.rolling(20).std()
mom_5d       = retornos_ars.rolling(5).sum()
mom_20d      = retornos_ars.rolling(20).sum()

# Para VIST y MELI calcular retornos en USD
retornos_adr = {}
for ticker in PANEL_ADR:
    col = f'{ticker}_USD'
    retornos_adr[f'ret_{ticker}']   = precios_ars[col].pct_change()
    retornos_adr[f'vol_{ticker}']   = precios_ars[col].pct_change().rolling(20).std()
    retornos_adr[f'mom5_{ticker}']  = precios_ars[col].pct_change().rolling(5).sum()
    retornos_adr[f'mom20_{ticker}'] = precios_ars[col].pct_change().rolling(20).sum()
retornos_adr_df = pd.DataFrame(retornos_adr, index=precios_ars.index)

retornos_ars.columns = [f'ret_{c}'   for c in acciones_locales]
vol_20d.columns      = [f'vol_{c}'   for c in acciones_locales]
mom_5d.columns       = [f'mom5_{c}'  for c in acciones_locales]
mom_20d.columns      = [f'mom20_{c}' for c in acciones_locales]

```

Fuente: elaboración propia.

Anexo 10

Función de optimización de Markowitz

El presente fragmento corresponde al script de la Etapa 2 y reproduce la función utilizada para resolver el problema de optimización de cartera. La función maximiza el ratio de Sharpe esperado bajo las restricciones de pesos no negativos, suma igual a uno y peso máximo por activo definido por la constante PESO_MAX. Esta es la capa de optimización que recibe como insumo las predicciones de los tres modelos de machine learning para construir la cartera final de cada jornada

```
# ================================================================
# PASO 2 — FUNCIÓN DE MARKOWITZ
# ================================================================
# Maximiza Sharpe usando solo datos históricos disponibles hasta
# el día T (ventana deslizante de VENTANA_COV días hacia atrás).
# Nunca usa datos futuros.

def optimizar_markowitz(retornos_esperados, cov_matrix, peso_max=PESO_MAX):
    n   = len(retornos_esperados)
    ret = np.array(retornos_esperados)
    cov = np.array(cov_matrix)

    def neg_sharpe(w):
        pr = np.dot(w, ret)
        pv = np.sqrt(np.dot(w.T, np.dot(cov, w)))
        return -pr / (pv + 1e-8)

    constraints = [{'type': 'eq', 'fun': lambda w: np.sum(w) - 1}]
    bounds      = [(0, peso_max)] * n
    w0          = np.ones(n) / n

    result = minimize(neg_sharpe, w0, method='SLSQP',
                      bounds=bounds, constraints=constraints,
                      options={'maxiter': 1000, 'ftol': 1e-9})
    return result.x if result.success else np.ones(n) / n

```

Fuente: elaboración propia.

Anexo 11

Random Forest con control de lookahead bias

El presente fragmento corresponde al script de la Etapa 2 y reproduce el entrenamiento del modelo Random Forest. Se entrena un clasificador independiente por cada acción del panel para estimar la probabilidad de retorno positivo al día siguiente. El uso del operador shift(-1) dentro del período de entrenamiento, junto con la separación estricta train/test fijada en mayo de 2019, garantiza que el modelo no acceda a información futura al momento de generar predicciones.

```
scaler_rf    = StandardScaler()
X_train_sc   = scaler_rf.fit_transform(train[COLS_FEATURES].fillna(0))
X_test_sc    = scaler_rf.transform(test[COLS_FEATURES].fillna(0))

modelos_rf = {}
for accion in ACCIONES:
    col_ret = f'ret_{accion}'
    if col_ret not in features.columns:
        continue
    # Target: 1 si el retorno del DÍA SIGUIENTE es positivo
    # shift(-1) en train: el día T predice T+1, que sigue siendo train
    y_train = (train[col_ret].shift(-1) > 0).astype(int).fillna(0)
    rf = RandomForestClassifier(
        n_estimators=200, max_depth=5, min_samples_leaf=10,
        class_weight='balanced', random_state=SEED_CORRIDA, n_jobs=-1)
    rf.fit(X_train_sc, y_train)
    modelos_rf[accion] = rf
```

Fuente: elaboración propia.

Anexo 12

Gradient Boosting con regularización (XGBoost)

El presente fragmento corresponde al script de la Etapa 2 y reproduce el entrenamiento del modelo XGBoost. Al igual que en el caso de Random Forest, se entrena un clasificador binario independiente por cada acción del panel para estimar la probabilidad de retorno positivo al día siguiente. La configuración adoptada incorpora dos mecanismos de regularización: el submuestreo aleatorio de filas y columnas durante el entrenamiento (subsample y colsample_bytree del 80%) y una tasa de aprendizaje moderada de 0,05, que en conjunto reducen el riesgo de sobreajuste a las particularidades del período de entrenamiento.

```
modelos_xgb = {}
for accion in ACCIONES:
    col_ret = f'ret_{accion}'
    if col_ret not in features.columns:
        continue
    y_train = (train[col_ret].shift(-1) > 0).astype(int).fillna(0).values
    y_test  = (test[col_ret].shift(-1) > 0).astype(int).fillna(0).values
    model = XGBClassifier(
        n_estimators=200, max_depth=4, learning_rate=0.05,
        subsample=0.8, colsample_bytree=0.8, eval_metric='logloss',
        random_state=SEED_CORRIDA, n_jobs=-1, verbosity=0)
    model.fit(X_train_sc, y_train,
              eval_set=[(X_test_sc, y_test)], verbose=False)
    modelos_xgb[accion] = model
```

Fuente: elaboración propia.

Anexo 13

Arquitectura LSTM con control de secuencias temporales

El presente fragmento corresponde al script de la Etapa 2 y reproduce la implementación de la red neuronal recurrente con arquitectura LSTM. Se entrena un modelo independiente por acción y se incorpora la función make_seq que construye las secuencias temporales respetando la lógica de que la secuencia del paso k utiliza únicamente los datos de los VENTANA_LSTM días previos a k. La arquitectura incluye dos capas recurrentes con dropout, una capa densa intermedia y mecanismo de detención temprana sobre el error de validación.

```python
def make_seq(X, y, ventana):
    """
    Construye secuencias temporales.
    Secuencia k: X[k-ventana : k] → predice y[k]
    Solo usa información anterior al instante k.
    """
    Xs, ys = [], []
    for k in range(ventana, len(X)):
        Xs.append(X[k - ventana : k])
        ys.append(y[k])
    return np.array(Xs), np.array(ys)

modelos_lstm = {}
for accion in ACCIONES:
    col_ret = f'ret_{accion}'
    if col_ret not in features.columns:
        continue

    y_all = (features[col_ret].shift(-1) > 0).astype(int).fillna(0).values
    X_seq, y_seq = make_seq(X_arr_sc, y_all, VENTANA_LSTM)

    # Corte: las primeras corte_seq secuencias son de train
    corte_seq = idx_corte - VENTANA_LSTM
    X_tr, y_tr = X_seq[:corte_seq], y_seq[:corte_seq]
    X_te, y_te = X_seq[corte_seq:], y_seq[corte_seq:]

    tf.random.set_seed(SEED_CORRIDA)
    modelo = Sequential([
        LSTM(64, return_sequences=True,
             input_shape=(VENTANA_LSTM, X_arr_sc.shape[1])),
        Dropout(0.2),
        LSTM(32),
        Dropout(0.2),
        Dense(16, activation='relu'),
        Dense(1,  activation='sigmoid'),
    ])
    modelo.compile(optimizer='adam',
                   loss='binary_crossentropy',
                   metrics=['accuracy'])
    es = EarlyStopping(monitor='val_loss', patience=8,
                       restore_best_weights=True, verbose=0)
    modelo.fit(X_tr, y_tr, epochs=60, batch_size=32,
               validation_split=0.2, callbacks=[es], verbose=0)
    modelos_lstm[accion] = modelo
```

Fuente: elaboración propia.

Anexo 14

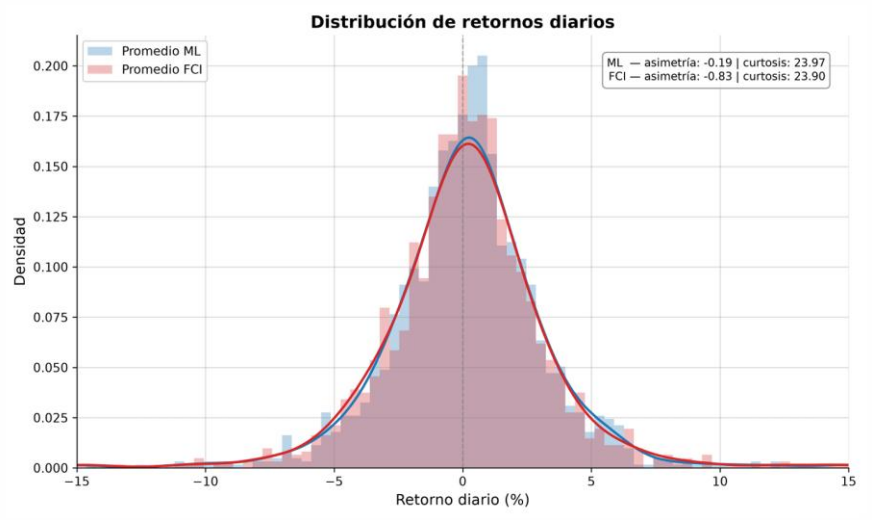
Simulación del fondo con rebalanceo diario

El presente fragmento corresponde al script de la Etapa 2 y reproduce la función que simula la evolución diaria del fondo. La función itera día por día sobre el período de evaluación valorizando las posiciones con los precios efectivamente observados, calculando el retorno diario y ajustando las cantidades de cada activo al peso objetivo determinado por el modelo el día anterior. El rebalanceo se implementa de manera instantánea y sin costos, y la valorización utiliza los precios del día posterior al de la decisión de pesos, evitando cualquier filtración temporal de información.

```python
def simular_fondo(pesos_df, precios_df, acciones, capital_inicial):
    """
    Simulación diaria básica sin costos ni restricciones.

    Día T:
      1. Valorizar posiciones con precios del día T
      2. Calcular retorno diario
      3. Ajustar cantidades exactamente a los pesos objetivo
         (rebalanceo perfecto e instantáneo)

    El fondo se valoriza con los precios del día SIGUIENTE
    al de la decisión de pesos — no hay lookahead.
    """
    idx      = pesos_df.index.intersection(precios_df.index)
    pesos_t  = pesos_df.loc[idx][acciones].fillna(0)
    precios  = precios_df.loc[idx][acciones]
    n_acc    = len(acciones)

    registros   = []
    cant_actual = np.zeros(n_acc)
    valor_fondo = float(capital_inicial)

    for i, fecha in enumerate(idx):
        precio_hoy = precios.loc[fecha].values.astype(float)
        peso_obj   = pesos_t.loc[fecha].values.astype(float)

        # Normalizar pesos (suman 1)
        sp = peso_obj.sum()
        if sp > 1e-8:
            peso_obj = peso_obj / sp
        else:
            peso_obj = np.zeros(n_acc)

        if i == 0:
            # Día inicial: comprar según pesos con el capital completo
            cant_actual = np.where(
                precio_hoy > 0,
                (valor_fondo * peso_obj) / precio_hoy,
                0.0
            )
            ret_diario = 0.0

        else:
            # 1. Valorizar posiciones a precios de hoy
            valor_pos  = np.nansum(
                cant_actual * np.where(precio_hoy > 0, precio_hoy, 0.0))
            ret_diario = ((valor_pos - valor_fondo) / valor_fondo
                          if valor_fondo > 0 else 0.0)
            valor_fondo = valor_pos

            if valor_fondo <= 0:
                registros.append({
                    'Fecha':           fecha,
                    'Valor_Fondo_ARS': 0.0,
                    'Ret_Diario_%':    -100.0,
                    'Ret_Acum_%':      -100.0,
                    'Periodo':         asignar_periodo(fecha),
                    **{f'Peso_{acciones[j]}': 0.0 for j in range(n_acc)},
                })
                continue

            # 2. Rebalancear: ajustar cantidades a los nuevos pesos
            cant_actual = np.where(
                precio_hoy > 0,
                (valor_fondo * peso_obj) / precio_hoy,
                0.0
            )
```

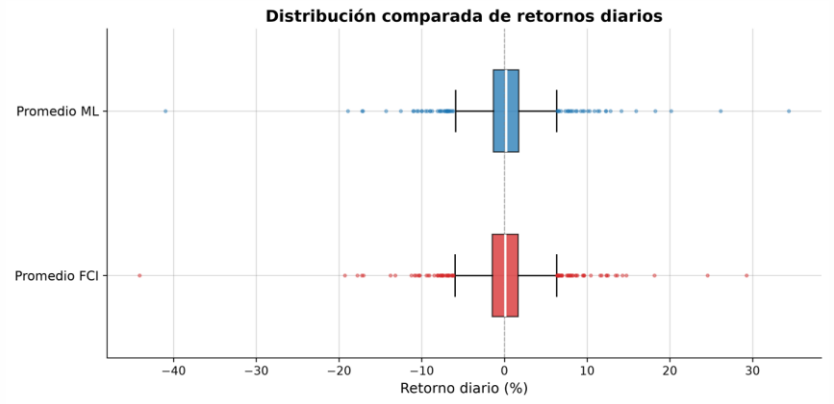
Fuente: elaboración propia.

Anexo 15



Distribución de los retornos diarios del promedio ML y el promedio FCI durante el período completo. Las barras corresponden al histograma de frecuencias y las curvas superpuestas a estimaciones de densidad. El recuadro reporta los valores de asimetría y curtosis exceso para cada serie. Fuente: elaboración propia.

Anexo 16



Distribución comparada de los retornos diarios del promedio ML y el promedio FCI mediante diagrama de cajas y bigotes. Las cajas representan el rango intercuartílico, la línea central indica la mediana, los bigotes se extienden hasta 1,5 veces el rango intercuartílico y los puntos corresponden a observaciones extremas. Fuente: elaboración propia.

Bibliografía

Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2015). The probability of backtest overfitting. *Journal of Computational Finance*, *20*(4), 39–69.

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, *116*(1), 261–292.

Bekaert, G., & Harvey, C. R. (2002). Research in emerging markets finance: Looking to the future. *Emerging Markets Review*, *3*(4), 429–448.

Bikhchandani, S., & Sharma, S. (2000). Herd behavior in financial markets. *IMF Staff Papers*, *47*(3), 279–310.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.*

Chapman & Hall.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417.

Fama, E. F., & French, K. R. (2010). *Luck versus skill in the cross-section of mutual fund returns. The Journal of Finance,* 65(5), 1915–1947.

Grobys, K., Kolari, J. W., & Niang, J. (2022). Man versus machine: On artificial intelligence and hedge funds performance. *Applied Economics*, *54*(40), 4632–4646.

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, *70*(3), 393–408.

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

Harvey, C. R. (1995). Predictable risk and returns in emerging markets. *The Review of Financial Studies*, *8*(3), 773–816.

Harvey, C. R., Rattray, S., Sinclair, A., & Van Hemert, O. (2017). Man vs. machine: Comparing discretionary and systematic hedge fund performance. *The Journal of Portfolio Management*, *43*(4), 55–69.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65–91.

Jobson, J. D., & Korkie, B. M. (1981). *Performance hypothesis testing*

with the Sharpe and Treynor measures. The Journal of Finance, 36(4),

889–908.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689–702.

López de Prado, M. (2018). *Advances in financial machine learning*. John Wiley & Sons.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35.

Memmel, C. (2003). *Performance hypothesis testing with the Sharpe ratio.*

Finance Letters, 1(1), 21–23.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Puetz, A., & Ruenzi, S. (2011). Overconfidence among professional investors: Evidence from mutual fund managers. *Journal of Business Finance & Accounting*, *38*(5–6), 684–712.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). *Dataset shift in machine learning*. MIT Press.

Sautu, R. (2010). *Manual de metodología: Construcción del marco teórico,*

formulación de los objetivos y elección de la metodología. Prometeo Libros.

Shefrin, H. (2000). *Beyond greed and fear: Understanding behavioral finance and the psychology of investing*. Harvard Business School Press.

Shefrin, H., & Statman, M. (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. *The Journal of Finance*, *40*(3), 777–790.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *1*(1), 39–60.

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing Science*, *4*(3), 199–214.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Wilcoxon, F. (1945). *Individual comparisons by ranking methods.*

Biometrics Bulletin, 1(6), 80–83.