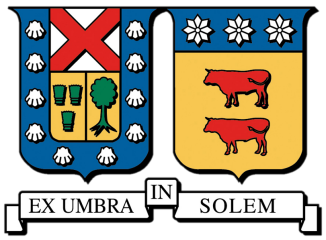


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE MATEMÁTICA
VALPARAÍSO - CHILE

Risk-Aware Portfolio Optimization via Reinforcement Learning with Expected Shortfall

---

Tesis presentada por:

Rodrigo Sebastián Serrano Pérez

Como requisito para optar al título profesional de Ingeniero Civil Matemático y el grado académico de Magíster en Ciencias de la Ingeniería Industrial

Director de Tesis:

Dr. Werner Kristjanpoller

Profesor Correferente:

Dr. Francisco Cuevas

Noviembre, 2025

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☐ Memoria o trabajo de título ☒ Tesis de Postgrado

Título del trabajo: Risk-Aware Portfolio Optimization via Reinforcement Learning with Expected Shortfall

Nombre del candidato(a): Rodrigo Sebastián Serrano Pérez

Carrera / Grado: Magíster en Ciencias de la Ingeniería Industrial

Campus: Casa Central **Departamento:** Industrias

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Werner David Kristjanpoller Rodriguez, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

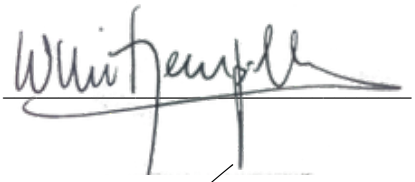
☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

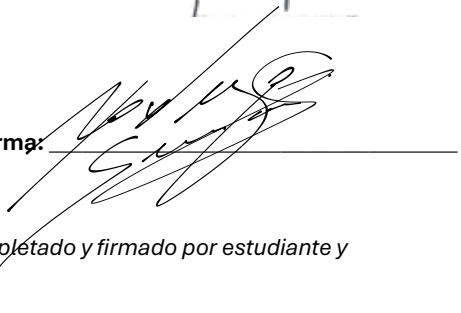
Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 24/11/2025 **Firma:** 

Estudiante o Candidato(a):

Fecha: 24/11/25 **Firma:** 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

TÍTULO DE LA TESIS:

RISK-AWARE PORTFOLIO OPTIMIZATION VIA REINFORCEMENT LEARNING WITH EXPECTED SHORTFALL.

AUTOR:

RODRIGO SERRANO PÉREZ

TRABAJO DE TESIS, presentado en cumplimiento parcial de los requisitos para el título de Ingeniero Civil Matemático y el grado académico de Magíster en Ciencias de la Ingeniería Industrial de la Universidad Técnica Federico Santa María.

COMISIÓN DE TESIS:

Integrantes	Firma
Dr. Werner Kristjanpoller Universidad Técnica Federico Santa María, Chile.	_____
Dr. Francisco Cuevas Universidad Técnica Federico Santa María, Chile.	_____
Dr. Felipe Escudero Universidad Técnica Federico Santa María, Chile.	_____

Valparaíso, Noviembre 2025.

RESUMEN

Esta tesis estudia series temporales de retornos logarítmicos de activos financieros líquidos con datos diarios entre **junio de 2018 y agosto de 2025**. El universo considera criptomonedas, ETF de renta variable estadounidense y bonos del Tesoro de larga duración, y acciones transadas en el mercado de estados unidos. Se incluyen propiedades empíricas relevantes para la gestión del riesgo (Varianza, Expected Shortfall, etc.), así como patrones de comovimiento entre clases de activos, destacando la **correlación negativa promedio** entre SPY y TLT y la **heterogeneidad sectorial** entre AAPL y JPM.

Sobre esta base descriptiva, se integra *Reinforcement Learning* a la Teoría de Portafolios. Se construye un entorno donde el agente observa ventanas deslizantes de retornos ($w \times N$) aplanadas y toma acciones discretas por activo $\{-1, 0, +1\}$. El entrenamiento se realiza con algoritmo **PPO** (Proximal Policy Optimization) y **recompensa episódica basada en el índice de Sharpe anualizado**, alineando el aprendizaje con un criterio de desempeño ajustado por volatilidad. Para su evaluación fuera de muestra, la señal direccional del agente se calibra mediante *binning* que mapea el *score* a $(\widehat{\mu}, \widehat{\sigma})$, y la decisión operativa se implementa con un **umbral dinámico** $\tau_t(k) = k\,\tilde{\sigma}_t$, donde $\tilde{\sigma}_t$ es una medida de riesgo **ex–ante** basada en **Expected Shortfall** con esquema EWMA de doble velocidad y corrección *ex–ante* para evitar *look–ahead*. El parámetro k se selecciona por **backtesting** en validación bajo criterios complementarios (Sharpe, ES, retorno acumulado y varianza) y luego se evalúa en un conjunto de *test*.

Los resultados descriptivos confirman el carácter leptocúrtico de los retornos en cripto, la capacidad de diversificación de TLT frente a SPY y la diversidad sectorial entre AAPL y JPM. El marco desarrollado es **reproducible**, pues se documentan fuente de datos (Yahoo Finance), precios ajustados por dividendos y *splits*, alineación de calendarios, partición estrictamente temporal y semillas de inicialización, además de controles para evitar fuga de información. En conjunto, el aporte es mostrar un procedimiento claro para estudiar estrategias de portafolio que combinan aprendizaje por refuerzo, calibración fuera de muestra y control del riesgo de cola, ofreciendo una base para extensiones futuras en selección de señales, modelos de riesgo y universos de activos.

AGRADECIMIENTOS

Primeramente, doy gracias a Dios por guiarme y darme la fortaleza en este camino.

También, a mi familia: mi padre, Jorge Serrano, quien ha sido un pilar fundamental y me ha brindado su apoyo incondicional y sabias palabras de aliento; y a mi madre, Susana Pérez, que me guía desde el cielo y cuya memoria me ha impulsado a no rendirme nunca y siempre querer ser mejor persona. A Sharay, mi pareja, quien ha sido mi compañera de vida en este proceso y ha estado a mi lado en cada momento, generando recuerdos imborrables.

A mi mejor amigo, Maximiliano Nuñez, que ha sido el hermano que nunca tuve; y a mis mascotas: Polo, Raysa y Chiqui, fieles compañeros en aquellas frías noches de desvelo para aquellas determinantes pruebas de asignaturas complejas.

A mis profesores guías, Werner y Francisco, quienes me ayudaron a encontrar el tema que me ha mantenido encantado en este camino, y que contribuyeron a formar una base sólida para mi desarrollo profesional.

Finalmente, a todos quienes, de alguna u otra forma, han aportado en este trabajo: muchas gracias.

Dedico este trabajo a mi madre, cuya ausencia física nunca ha impedido que su amor y enseñanza traspasen dimensiones y me llenen de fuerzas. Un beso y mil abrazos al cielo Mamá.

Índice general

Capítulo 1

Introducción

En el contexto actual de los mercados financieros, caracterizados por una alta volatilidad, interconexiones globales, y creciente disponibilidad de acceso a los datos, algunos métodos tradicionales de optimización de portafolios y de predicción han comenzado a mostrar algunas limitaciones significativas, como el hecho de ignorar que los errores de predicción afectan las decisiones finales. Pues, los modelos clásicos se separan principalmente en predicción de retornos y optimización de pesos. Algunas estrategias como puede ser la de asignación óptima de portafolio de la teoría del portafolio de Markowitz[1], basada en supuestos gaussianos sobre los retornos y correlaciones lineales entre los activos, resultan un tanto ineficientes frente a las complejidades y dinámicas de los mercados modernos. En especial, cuando se está buscando capturar riesgos más extremos (como crisis económicas) o adaptarse a perfiles de riesgo que sean heterogéneos entre inversores.

Por otro lado, se ha visto el gran avance que ha ido teniendo la inteligencia artificial y el aprendizaje automático. Esto ha permitido el desarrollo de nuevas herramientas que capturen de forma más eficiente relaciones no lineales o patrones subyacentes en los datos financieros. Pues, en general las reacciones de mercado a noticias o shocks son asimétricas, existiendo una reacción más errática a las malas noticias). Dentro de ellas, el *Reinforcement Learning* (RL) ha ido emergiendo como un enfoque prometedor en la necesidad de tomar decisiones secuenciales bajo incertidumbre, especialmente por la capacidad que tiene de aprender políticas

[1]Se le llama estrategia dado que en su implementación real, el modeldo define cómo se asigna el capital entre activos según sus riesgos y correlaciones. Parte de una estrategia estática.

óptimas en entornos secuenciales y dinámicos. De igual forma, el algoritmo *Proximal Policy Optimization* o simplemente PPO, ha demostrado buenos resultados en tareas relacionadas con la toma de decisiones complejas, siendo una implementación muy utilizada en finanzas, pues posee un método de actor-crítico, permitiendo juzgar aquellas decisiones tomadas por el RL, lo que le permite ser una buena opción para lo que significan las decisiones sobre un activo.

Esta tesis tiene como objetivo general desarrollar un modelo de optimización de portafolios basado en PPO, que permite ajustar los pesos asignados a distintos activos financieros con el fin de maximizar las ganancias ajustadas por riesgo. Para ello, se diseñó un entorno de inversión simulado que busca replicar condiciones reales de mercado, incorporando medidas de riesgo robustas como el *Expected Shortfall* (ES).

A lo largo del desarrollo de esta tesis se abordarán diversas fases metodológicas, desde la selección de activos, hasta la implementación del modelo PPO como agente capaz de aprender decisiones óptimas de inversión. Además, se incluirá una evaluación comparativa frente a otros métodos clásicos de optimización como Markowitz.

De esta forma, se puede posicionar la investigación en la intersección entre las finanzas cuantitativas y el aprendizaje profundo, con la finalidad de aportar un marco metodológico reproducible y flexible (adaptabilidad para distintos activos o perfiles de riesgo) para la gestión de portafolios bajo incertidumbre. Desarrollando un enfoque de adaptar la política de inversión al perfil de riesgo del inversionista, lo cual se operacionaliza a través de funciones de recompensa personalizadas.

Capítulo 2

Nociones Básicas

En finanzas, un **activo financiero** es un instrumento que representa un derecho contractual a recibir dinero u otro activo, ya sea en forma de propiedad (acciones) o deuda (bonos), entre otros. Estos activos se negocian en mercados financieros (de capitales o de valores), que pueden ser primarios (emisión) o secundarios (intercambio entre inversionistas). Su precio en el tiempo, definido como P_t, se determina por la interacción entre oferta y demanda, y refleja tanto información pública como expectativas del mercado.

Esta tesis, se centra en estudiar activos líquidos cuyo precio se observa en intervalos regulares de tiempo. Se considera tanto activos tradicionales (acciones, índices bursátiles, ETFs) como criptomonedas (que presentan mayor volatilidad y comportamientos de colas pesadas).

Una medida fundamental para cuantificar la evolución de un activo en el tiempo es su **retorno**, que captura la variación proporcional de su precio entre dos instantes. Aunque existen varias formas de calcularlo (Por ejemplo, mediante la variación porcentual o diferencia logarítmica), se prefiere el *retorno logarítmico*, ya que permite una representación aditiva en el tiempo y facilita el tratamiento estadístico, como se describe a continuación.

2.0.1. Métricas Financieras

Definición 1 Suponga que se tiene un portafolio de n activos financieros que actualmente tienen un valor de mercado de P_t^p, de forma que en $t+1$, el precio cambia a P_{t+1}^p. El **retorno logarítmico** o rendimiento logarítmico generado del portafolio, vendrá dado según:

$$R_{t+1}^p = \ln\left(\frac{P_{t+1}^p}{P_t^p}\right),$$

a diferencia del retorno porcentual, el retorno logarítmico posee la **propiedad de aditividad en el tiempo**. Esto significa que el retorno acumulado entre dos fechas cualesquiera t y $t+\Delta t$ (con $\Delta t > 0$) puede obtenerse simplemente como la suma de los retornos logarítmicos intermedios. Lo anterior, pues notar que:

$$\frac{P_{t+\Delta t}}{P_t} = \frac{P_{t+\Delta t}}{P_{t+\Delta t-1}} \cdot \frac{P_{t+\Delta t-1}}{P_{t+\Delta t-2}} \cdot \ \ldots \ \cdot \frac{P_{t+1}}{P_t} = \prod_{k=1}^{\Delta t} \frac{P_{t+k}}{P_{t+k-1}},$$

y usando de que $\ln(a \cdot b) = \ln(a) + \ln(b)$, se tiene que:

$$\begin{aligned} \ln\left(\frac{P_{t+\Delta t}}{P_t}\right) &= \ln\left(\prod_{k=1}^{\Delta t} \frac{P_{t+k}}{P_{t+k-1}}\right) \\ &= \sum_{k=1}^{\Delta t} \ln\left(\frac{P_{t+k}}{P_{t+k-1}}\right), \end{aligned}$$

lo cual permite descomponer y recomponer retornos de forma lineal, facilitando el análisis estadístico. Bajo algunas suposiciones, el portafolio de inversión tiene la siguiente forma:

$$\textit{portafolio} = [A_1 \cdot a_1, A_2 \cdot a_2, ..., A_n \cdot a_n],$$

en donde A_i es el número de posiciones del activo a_i. Entonces, el precio en t del portafolio, está dado según:

$$P_t^p = A_1 \cdot P_t^{a_1} + A_2 \cdot P_t^{a_2} + ... + A_n \cdot P_t^{a_n}.$$

Luego, el retorno del portafolio se podría aproximar de forma lineal mediante:

$$R_{t+1}^p = A_1 \cdot R_{t+1}^{a_1} + A_2 \cdot R_{t+1}^{a_2} + ... + A_n \cdot R_{t+1}^{a_n} = \sum_{i=1}^{n} A_i \cdot R_{t+1}^{a_i}.$$

El retorno del portafolio, tiene un riesgo intrínseco. En el contexto de la teoría moderna de portafolios Markowitz (1952), el riesgo se modela habitualmente como

la *varianza* de los retornos del portafolio, denotada por σ_p^2.

Un **problema de optimización de portafolios** consiste en asignar pesos $w = (w_1, w_2, \ldots, w_n)$ a cada activo de forma que se alcance un objetivo (por ejemplo, maximizar retorno o minimizar riesgo) sujeto a ciertas restricciones (presupuesto, límites regulatorios, tolerancia al riesgo, etc.).

Bajo la consideración de que los activos poseen una correlación entre ellos, las cuales están en Σ, la cual es una matriz cuadrada de $n \times n$ que contiene las covarianzas entre los retornos de pares de activos $\sigma_{i,j} = \text{Cov}(R_i, R_j)$, se tendría entonces que el riesgo del portafolio está dado según:

$$\sigma_p^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij} = w^T \Sigma w$$

En este marco, la formulación clásica propuesta por Markowitz (1952) plantea:

- Maximizar el retorno esperado del portafolio sujeto a un nivel de riesgo máximo permitido.
- Minimizar el riesgo sujeto a un nivel mínimo de retorno esperado.

Una de sus formas más comunes es:

$$\min_w \; w^\top \Sigma w \quad \text{sujeto a} \quad \mathbb{E}(R_p) = \sum_i w_i \cdot \mathbb{E}(R_i), \quad \sum_i w_i = 1.$$

Diversos estudios han demostrado que los retornos financieros presentan características empíricas como asimetría, curtosis excesiva y colas pesadas. Las primeras evidencias que marcaron el análisis de series financieras vinieron de Mandelbrot (1963), en donde observó que los retornos de los activos financieros, como los del algodón, no poseían una distribución normal. Que en cambio presentaban colas más gruesas, implicando una probabilidad mayor de eventos extremos, cuestionando así el uso de la distribución normal en la modelación del riesgo y valorización de activos. Su trabajo también incluye un punto bastante importante, el cual es la evidencia de que los movimientos o volatilidades de los precios poseen patrones como autocorrelación o agrupamiento, que se mantienen similares a distintas temporalidades, lo que fue un anticipo de la necesidad del uso de modelos que

incorporen volatilidades variables y colas pesadas.

Con ello entonces, Mandelbrot produjo el punto de quiebre con el supuesto de normalidad, poniendo en duda la visión clásica de los retornos financieros con varianza constante. Naciendo así la necesidad de encontrar un modelo que entienda la varianza como una variable temporal.

Durante la década del 70 y los principios del 80, aún en ausencia de los modelos requeridos según lo mencionado en el párrafo anterior, se utilizaban modelos de series temporales ARIMA o de media móvil, descubiertos por Box and Jenkins (1970), los cuales asumían **homocedasticidad**, es decir, varianza constante en el tiempo. Permaneciendo así el desafío de poder encontrar un modelo adecuado según los resultados obtenidos por Mandelbrot.

Engle (1982), introduce el modelo *ARCH* (Autoregressive Conditional Heteroskedasticity), modelo en el cual la varianza condicional en t dependía linealmente de los errores al cuadrado de los periodos anteriores. Su formulación se define de la siguiente manera:

Definición 2 Sea $\{R_t\}$ una serie temporal definida como

$$R_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t \mid \mathcal{F}_{t-1} \sim (0, \sigma_t^2),$$

donde $\mathcal{F}_{t-1}$ es la información disponible hasta el periodo $t-1$ y σ_t^2 es la varianza condicional del error[1]. Se dice que $\{R_t\}$ sigue un modelo *Autoregressive Conditional Heteroskedasticity* de orden q (ARCH(q)), si la varianza condicional del error está dada por

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2,$$

con $\alpha_0 > 0$ y $\alpha_i \geq 0$ para todo $i = 1, \ldots, q$. Posteriormente, Bollerslev (1986) generalizó lo propuesto por Engle, mediante el modelo *GARCH* o también llamado *Generalized ARCH*, el cual sigue la definición:

Definición 3 Sea $\{R_t\}$ una serie temporal con media condicional $\mu_t = \mathbb{E}(R_t \mid \mathcal{F}_{t-1})$,

[1]Se puede interpretar como la volatilidad esperada en t o el tamaño del shock esperado en el retorno, condicionado a la información pasada.

donde $\mathcal{F}_{t-1}$ representa la información disponible hasta el tiempo $t-1$. El modelo **GARCH** (*Generalized Autoregressive Conditional Heteroskedasticity*) se define como:

$$R_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \mid \mathcal{F}_{t-1} \sim (0, \sigma_t^2),$$
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2,$$

donde $\alpha_0 > 0$, $\alpha_i \geq 0$ y $\beta_j \geq 0$ para todo i, j.

En el modelo *GARCH*, el término σ_t^2 depende tanto de los errores pasados como de los valores previos de la propia varianza, logrando así capturar la persistencia temporal y agrupamiento de volatilidad descrito inicialmente por Mandelbrot.

A pesar de este desarrollo, existía la necesidad institucional (por parte de la reguladora de la banca, como la CMF en Chile) de medir y controlar el **riesgo de mercado** de forma integral para grandes carteras de activos. Una respuesta, fue el *Value at Risk*, que aunque sus fundamentos teóricos están más ligados a la estadística y la teoría de portafolios propuesta por Markowitz (1952), su principal formulación y uso se atribuye a J.P. Morgan, quienes en 1996 desarrollaron y publicaron RiskMetrics (RiskMetrics Group (1996)), que era una metodología estándar para el cálculo del VaR para carteras, lo que fue un gran hito para la gestión del riesgo financiero, tanto para bancos como para traders y/o inversionistas. El *VaR*, se define a continuación:

Definición 4 Sea R_X el retorno de un activo X, y L_X la pérdida (o retorno negativo) del activo X, se define el Valor en Riesgo (o Value at Risk) según:

$$VaR_\alpha(X) = -\inf\{c : \mathbf{P}(R_X \geq -c) \geq \alpha\} = -\inf\{c : \mathbf{P}(L_X \leq c) \geq \alpha\},$$

donde $\mathbf{P}(\cdot)$ denota la medida de probabilidad, que básicamente se refiere a el nivel de pérdida que no será superado con un $\alpha\,\%$ de confianza. El *VaR* ha sido una métrica **ampliamente** utilizada por reguladores financieros hasta el día de hoy (Jorion (2006))

No obstante, años posteriores (para finales de la década del 90), comenzaron a encontrar evidencias de limitaciones del VaR_α. Artzner et al. (1999) encontró que el principal problema, era de que el VaR_α no captura las pérdidas más allá del

umbral de confianza, lo que hace omitir los eventos de cola. En segundo lugar, que no cumple propiedades axiomáticas, en particular, la subaditividad. Lo que implica que el riesgo total de una cartera podría, en ciertos escenarios, ser mayor que la suma de los riesgos de los activos por sí solos. Artzner, propuso una serie de propiedades para así definir lo que llamarían como **Medida de Riesgo Coherente en el sentido Artzner**. Estas se presentan en la siguiente definición:

Definición 5 Sea $(\Omega, \mathcal{F}, \mathbf{P})$ un espacio de probabilidad, R_X el retorno de una inversión X y L_X la pérdida (o retorno negativo) de la inversión X, y sea también $\mathcal{L}$ el conjunto de variables aleatorias de *pérdida* asociada a X. Una **medida de riesgo** es una aplicación $\rho : \mathcal{L} \to \mathbb{R}$. Se dice que ρ es **coherente** en el sentido de Artzner si satisface, para todo $X, Y \in \mathcal{L}$, $m \in \mathbb{R}$ y $\lambda > 0$, las siguientes propiedades:

1. **Monotonía:** Si $X \leq Y$ casi seguramente, entonces $\rho(X) \leq \rho(Y)$,
2. **Invarianza traslacional:** $\rho(X + m) = \rho(X) + m$, [2]
3. **Homogeneidad positiva:** $\rho(\lambda X) = \lambda\, \rho(X)$,
4. **Subaditividad:** $\rho(X + Y) \leq \rho(X) + \rho(Y)$.

La subaditividad formaliza el principio de diversificación, el riesgo de una cartera de activos no debe exceder la suma de los riesgos de cada activo.

Aunque el VaR_α si satisface la monotonía, invarianza traslacional y homogeneidad positiva, **no es**, en general, **subaditivo**. Por lo que, no sería una medida de riesgo coherente en el sentido Artzner. La demostración de aquello, se encuentra en el Apéndice (7.2).

El VaR puede resultar subaditivo bajo supuestos restrictivos (p. ej., familias elípticas como la normal multivariante o bajo comonotonicidad), pero *no* lo es en general. Finalmente, Artzner realizó propuestas de métricas como el *Tail Conditional Expectation* o también llamado *TailVaR*, como una alternativa que si fuese coherente. La definición exacta que ellos entregan, es la siguiente:

[2] Añadir efectivo libre de riesgo m aumenta la cifra de pérdidas en m.

Definición 6 Sea R_X el retorno de la inversión X y r el rendimiento de referencia, se define el *Tail Conditional Expectation* como:

$$TCE_\alpha(X) = -\mathbb{E}[R_X/r \mid R_X/r \leq VaR_\alpha(X)]$$

Más adelante, Rockafellar and Uryasev (2000) propuso la utilización de una nueva métrica de riesgo, el **Conditional VaR** o simplemente (*CVaR*), que derivaba a partir de la definición del *TailVaR* hecha por Artzner et al. (1999). En este ámbito, el *CVaR* se define como:

Definición 7 Sea L_X una variable aleatoria que representa la *pérdida* asociada a una inversión X, y sea $\alpha \in (0,1)$ el nivel de confianza. El *Conditional Value-at-Risk* (CVaR) al nivel α se define como:

$$CVaR_\alpha(X) = \min_{c\in\mathbb{R}} \left\{ c + \frac{1}{1-\alpha}\,\mathbb{E}\big[(L_X - c)^+\big] \right\},$$

donde $(x)^+ = \max\{x, 0\}$.

Posteriormente, en paralelo a la búsqueda por encontrar una medida de riesgo coherente, Cont (2001) reforzó las observaciones empíricas realizadas por Mandelbrot respecto a los retornos de activos (por ejemplo, la curtosis elevada[3]), y definió lo que llamó como "hechos estilizados" que son patrones recurrentes que se mantienen estables a lo largo del tiempo. Cont descubrió ausencias de evidencias significativas respecto a que el retorno de hoy dependía linealmente del retorno de ayer, hecho definido por ausencia de autocorrelación lineal. También descubrió otras características los cuales dieron paso al uso de distintas herramientas para el modelamiento de series financieras, en su mismo artículo (Cont (2001)) detalla respecto a efectos de volumen y liquidez en la estructura de mercado, en particular, de que la volatilidad tiende a aumentar en períodos donde también aumentan el volumen de transacciones y la falta de liquidez, lo cual era una explicación del agrupamiento de volatilidades descubierto por Mandelbrot. Con ello, Cont logró probar mediante pruebas estadísticas sobre autocorrelación y heterocedasticidad, que muchos de los modelos existentes hasta esa fecha (como Black-Scholes, modelos AR o ARIMA) no reproducen estos "hechos estilizados", justificando así el uso de

[3]La curtosis elevada afecta a medidas de riesgo tradicionales como la varianza al subestimar las pérdidas extremas

modelos alternativos como los *GARCH*, evidenciando que el uso tradicional de la varianza basada en el supuesto de normalidad, resultaba insuficiente para la cuantificación del riesgo.

Siguiendo con las métricas de riesgo, resultó muy importante el artículo publicado por Artzner et al. (1999), pues deja en evidencia de que el *VaR* no era coherente en su sentido. A su vez, Rockafellar and Uryasev (2000) daban los primeros pasos respecto a una nueva métrica de riesgo para la gestión del riesgo extremo. La cual, nacía a partir de la definición del *TailVaR* la cual, según Artzner, **si** era una medida coherente. No fue hasta el año 2002 en donde Acerbi and Tasche (2002) publicaron un artículo consolidando todas las definiciones de *CVaR, TailVaR, etc.* y definieron lo que se conoce como **Expected Shortfall**, la cual demostraron que era una medida coherente en el sentido de Artzner. El *ES* se definía según:

Definición 8 Sea R_X el retorno de una inversión X y L_X la pérdida, se define el *Expected Shortfall* como:

$$ES_\alpha(X) = -\frac{1}{1-\alpha}\mathbb{E}[R_X \mathbf{1}_{\{R_X \leq VaR_\alpha(X)\}}],$$

esta definición es general e incluye distribuciones no continuas. El $ES_\alpha(X)$, busca calcular la pérdida esperada dado que se excedió el VaR. En otras palabras, medir cual podría ser la pérdida promedio en el $(1-\alpha)\,\%$ de los casos.

En el mismo artículo de Acerbi and Tasche (2002), detallan la demostración de que el *ES* es coherente en el sentido de Artzner. No se entrará en detalle respecto a la demostración de la coherencia del ES en el sentido de Artzner para las propiedades de homogeneidad, invarianza y monotonicidad. Dicha demostración está disponible en su artículo Acerbi and Tasche (2002). Pero en esencia, la homogeneidad resulta a partir de la linealidad de la esperanza, la invarianza por traslación de que $\mathbb{E}[c] = c$ para $c \in \mathbb{R}$ una constante, mientras que la monotonicidad resulta de que si $X - Y \leq 0$, entonces $\mathbb{E}[X - Y] \leq \mathbb{E}[0] = 0 \implies \mathbb{E}[X] \leq \mathbb{E}[Y]$. La demostración de la subaditividad del Expected Shortfall, se puede revisar en el apéndice 7.2.

Más tarde, Chu et al. (2017) extendieron el trabajo realizado por Cont (2001) respecto a la evidencia de los hechos estilizados para un tipo de activo particularmente muy volátil: Las criptomonedas. En su estudio, mostraron que Bitcoin posee

propiedades consistentes con las observadas por Cont (colas pesadas, asimetría, volatilidad agrupada) reforzando así la idea del uso de métricas de riesgo coherentes con distribuciones no normales. Asimismo, evaluaron la capacidad del **VaR** para la estimación del riesgo extremo en dichos activos, donde observaron que esta métrica **subestima las pérdidas** en los niveles de confianza más altos. En particular, notaron de que el *VaR*, al depender de un único cuantil de la distribución, no incluye la información respecto a la posible severidad de las pérdidas más allá del umbral, el cual podría verse acentuado dada la presencia de no normalidad o colas pesadas. Concluyendo así que, tanto la volatilidad tradicional como el *VaR* son insuficientes para la cuantificación del riesgo subyacente para activos con retornos más erráticos y dependientes del régimen del mercado.

Estas características implican que los retornos más alejados tienden a ocurrir con mayor frecuencia que la predicha por una distribución normal. En particular, las colas pesadas significan que la varianza, al promediar las desviaciones cuadráticas, tiende a subestimar la magnitud e impacto de estos eventos extremos. Este fenómeno no es exclusivo de las criptomonedas, sino que es común en muchos activos de alta volatilidad, donde los riesgos de cola pueden ser una parte significativa de las posibles pérdidas. Por lo anterior, para esta tesis se utiliza una definición distinta de riesgo a la de la volatilidad tradicional. La cual será basada en el *Expected Shortfall* (ES), que permite capturar de forma más precisa la pérdida media condicionada. Para dar una idea gráfica del *ES*, en la Figura 2.1, que representa los retornos de la criptomoneda *TRON* (o *TRX* según su ticket en Yahoo Finance), el VaR delimita una cota para las pérdidas con cierta probabilidad, mientras que el Expected Shortfall es el promedio de las peores pérdidas, delimitada por la línea continua roja vertical.

Una de las métricas más utilizadas para evaluar el rendimiento de un portafolio de inversión ajustado al riesgo es el **Ratio de Sharpe**, propuesto en Sharpe (1966). Este ratio mide el exceso de retorno frente a una alternativa libre de riesgo (o también llamada prima de riesgo) que un portafolio genera por unidad de riesgo asumido. En otras palabras, evalúa si los retornos adicionales obtenidos frente a una alternativa libre de riesgo, compensa o no la incertidumbre del activo. Su formulación es la siguiente:

Definición 9 Sea R_X el retorno de una inversión X y R_f la tasa de retorno libre de riesgo (por ejemplo, el rendimiento de bonos del Tesoro a corto plazo), el

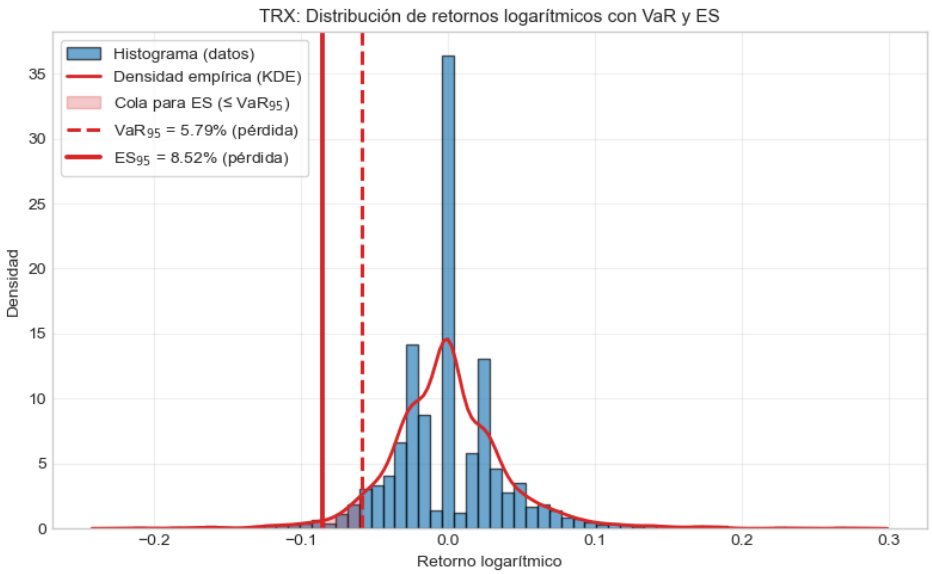


Figura 2.1. TRX: distribución empírica de retornos logarítmicos con $\text{VaR}_{5\,\%}$ y $\text{ES}_{5\,\%}$.

Ratio de Sharpe se define como:

$$SR_X = \frac{\overline{R_X} - R_f}{\sigma_X}, \tag{2.1}$$

donde $\overline{R_X}$ representa el promedio muestral de los retornos de la inversión X, R_f corresponde a la tasa libre de riesgo, y σ_X es la desviación estándar muestral de los retornos de la inversión X.

Si $SR_X > 0$ indica que la inversión X ha generado retornos superiores a la tasa libre de riesgo. Un valor más alto implica un mejor rendimiento ajustado por riesgo. En cambio, si $SR_X < 0$, entonces la inversión X ha rendido menos que la tasa libre de riesgo. Por lo que, las ganancias de la inversión X no han compensado el riesgo asumido, y una inversión en un activo libre de riesgo habría sido más rentable.

El Ratio de Sharpe es también muy útil para la comparativa de inversiones. Por ejemplo, si la inversión A tiene un retorno esperado mayor que la inversión B, pero también una volatilidad significativamente más alta, el Ratio de Sharpe clasifica cuál sería la mejor inversión en términos de retorno por riesgo asumido.

Es importante tener en cuenta que el ratio de Sharpe asume que el riesgo se mide adecuadamente por la desviación estándar, lo cual puede ser una limitación en distribuciones con asimetría y colas pesadas, como del caso de las criptomonedas.

En estos casos, medidas de riesgo alternativas como el *Expected Shortfall* poseen un mejor rendimiento.

Ajeno a lo anterior, por definición, el ratio de Sharpe se constriye a partir del promedio muestral de los retornos y de su varianza, los cuales pueden presentar dependencia temporal y heterocedasticidad condicional como se mencionó previo a la mención del *GARCH*. Por lo que, supuestos de independencia o igualdades en los errores en la varianza, no necesariamente se cumplen. En cuyo caso, el error estándar del ratio de Sharpe estaría sesgado, y entonces el contraste del exceso de retorno $H_0 : E[R_X] - R_f = 0$[4] requiere de algunas correcciones, especialmente si se quieren contrastar distintas estrategias de inversión para probar su superioridad. .

Para corregir el sesgo del SR_X, Newey and West (1987) propusieron el estimador **HAC** (*Heteroskedasticity and Autocorrelation Consistent*), el cual estima la varianza de largo plazo de la media o de los coeficientes estimados, inclusive cuando los errores presentan autocorrelación y varianza no constante, como es el caso en las series financieras. El estimador *HAC* se rige según la siguiente definición.

Definición 10. Sea $\{R_{X,t}\}_{t=1}^{T}$ la serie de retornos de una inversión X y R_{BH_X} el retorno generado por comprar el activo en $t = 1$ y mantener (no realizar un cambio de posición) hasta T. Sea también la regresión de contraste de exceso medio de retorno, definida como:

$$R_{X,t} - R_{BH_X} = \beta_0 + u_t,$$

donde $\beta_0 = \mathbb{E}[R_X - R_{BH_X}]$ es el exceso de retorno esperado, y u_t son los residuos de la regresión (las desviaciones respecto al exceso medio de retorno). Entonces, definimos el estimador **HAC** como:

$$\widehat{\Omega} = \Gamma_0 + \sum_{k=1}^{q} w_k \left(\Gamma_k + \Gamma_k^{\top}\right), \quad \text{con } \Gamma_k = \frac{1}{T}\sum_{t=k+1}^{T} u_t u_{t-k},$$

donde q es el número de rezagos considerados y w_k son pesos $w_k = 1 - k/(q+1)$.

Antes de Newey and West (1987), lo estimadores de varianza tipo σ^2/T se derivaban bajo el supuesto de que lo errores eran i.i.d., tal como se plantea en el modelo clásico

[4]Donde $\mathbb{E}[R_X]$ denota el valor esperado del retorno de la inversión X, es decir, su rendimiento medio teórico, que refleja la ganancia esperada por unidad de inversión

de regresión lineal de White (1980); Hamilton (1994). Sin embargo, bajo dependencia temporal y heterocedasticidad, dicho supuesto se viola, por lo que la varianza condicional subestima la verdadera dispersión del estimador. A partir de ello, Newey and West (1987), demostraron que la varianza del estimador de β_0 se aproxima por

$$\mathrm{Var}(\hat{\beta}_0) \approx \frac{\Omega}{T},$$

donde $\Omega = \sum_{k=-\infty}^{\infty} \Gamma_k$ es la varianza de largo plazo de los residuos. Al reemplazar Ω por su estimador muestral $\widehat{\Omega}$, se obtiene una estimación más consistente de la varianza del exceso medio de retorno.

Así, el error estándar del exceso de retorno medio corregido por heterocedasticidad y autocorrelación se obtiene como $\sqrt{\widehat{\Omega}/T}$, y el estadístico $t_{\mathrm{HAC}} = \hat{\beta}_0/\sqrt{\widehat{\Omega}/T}$ permite contrastar $H_0 : \mathbb{E}[R_X] - R_{BH_X} = 0$.

Newey and West (1987) discuten la elección de q y proponen utilizar:

$$q \;=\; \left\lfloor 4\left(\frac{T}{100}\right)^{2/9} \right\rfloor,$$

que equilibra el sesgo y la varianza del estimador en muestras finitas.

Para sintetizar, se tiene entonces que el estimador HAC corrige la inferencia del exceso de retorno medio al estimar la varianza de largo plazo de los residuos, incorporando la dependencia temporal mediante el truncamiento de q y pesos w_k. Sin embargo, cuando existen estructuras de dependencia no lineales o más complejas, una alternativa viable es la estimación de la misma varianza, pero de manera empírica mediante métodos de remuestreo dependiente.

Una alternativa para esta estimación empírica, son los procedimientos de *Moving Block Bootstrap*, que buscan replicar la correlación de los retornos sin la necesidad de una forma funcional empírica. Entre estos métodos, el MBB se ha consolidado como uno de los más utilizados para la construcción de intervalos de confianza del Ratio de Sharpe y contrastes de medias con dependencia, introducido en Künsch (1989).

En la gestión del riesgo de una inversión, la razón de utilizar el MBB no es meramente para estimar medias, sino que para evaluar la incertidumbre del rendimiento y del SR_X cuando el riesgo no puede medirse correctamente bajo supuestos clásicos i.i.d. El método de MBB fue propuesto en Künsch (1989) y posteriormente formalizado por Liu and Singh (1992). Este método remuestrea *bloques solapados* de la serie,

buscando preservar la estructura temporal dentro de cada bloque. El algoritmo, para nuestra aplicación, sigue el siguiente esquema:

Sea $\{R_t\}_{t=1}^{T}$ la serie de retornos de una inversión X. Fijada una longitud de bloque ℓ:

1. Construir los bloques solapados

$$B_j = (r_j, r_{j+1}, \ldots, r_{j+\ell-1}), \quad j = 1, \ldots, T - \ell + 1.$$

2. Sea $M = \lceil T/\ell \rceil$. Para cada réplica bootstrap $b = 1, \ldots, B$:

   a) Muestrear con reemplazo índices

$$J_1, \ldots, J_M \sim \text{Unif}\{1, \ldots, T - \ell + 1\}.$$

   b) Concatenar los bloques seleccionados $(B_{J_1}, B_{J_2}, \ldots, B_{J_M})$ y truncar la muestra resultante a longitud T para obtener la serie bootstrap

$$\{r_t^{*(b)}\}_{t=1}^{T}.$$

   c) Calcular la media bootstrap de los retornos:

$$\bar{r}^{*(b)} = \frac{1}{T} \sum_{t=1}^{T} r_t^{*(b)}.$$

3. El intervalo de confianza al $95\,\%$ para la media de retornos se obtiene con los percentiles $(2.5,\ 97.5)$ de la distribución empírica $\{\bar{r}^{*(b)}\}_{b=1}^{B}$:

$$IC_{95\,\%}(\bar{r}) = \left[\text{Percentil}_{2.5}(\{\bar{r}^{*(b)}\}),\ \text{Percentil}_{97.5}(\{\bar{r}^{*(b)}\})\right].$$

La dependencia de qué tan eficaz sea el método MBB proviene de la elección de la longitud de bloque ℓ. Pues, este parámetro determina el punto final de la estructura de dependencia temporal de los retornos preservada en cada remuestreo. En el caso en donde los bloques sean muy cortos, se pierde correlación serial, la cual es característica de los datos financieros. Mientras que si los bloques son muy largos, aumentan la varianza de las estimaciones del bootstrap y entonces se reduce la capacidad de replicabilidad de escenarios.

Por lo tanto, es importante realizar una correcta elección de ℓ, para buscar el equilibrio enre el sesgo por ruptura de la dependencia y la varianza por sobreajuste de dependencia. Para ello, la elección sigue una regla práctica estándar, basada en resultados asintóticos que minimizan el error cuadrático medio del estimador bootstrap de varianza. En particular, Hall et al. (1995); Lahiri (1999); Politis and White (2004) muestran que la longitud óptima crece a orden de $T^{1/3}$, por lo que se debiese escoger tal que

$$\ell \;=\; \text{máx}\{5,\, \lceil T^{1/3} \rceil\},$$

en donde 5 es una correción empírica propuesta por Lahiri (1999) de forma de evitar que en muestras pequeñas el tamaño del bloque sea corto.[5]

Para esta tesis en particular, se reporta el estimador puntual, el HAC y su p-valor para medir el exceso medio de retorno de una estrategia de inversión particular respecto a una estrategia pasiva como el Buy&Hold (comprar y no volver a operar o realizar cambio de posiciones). También, se reportan intervalos de confianza del SR_X vía MBB en presencia de dependencia temporal y heterocedasticidad.

Ya habiendo definido algunas métricas claves para la gestión u optimización de portafolios, tal como se menciona en el propósito de esta tesis, se busca modelar **toma de decisiones de inversión** bajo condiciones de incertidumbre, utilizando la información y métricas anteriormente mencionadas. Para ello, se emplea un modelo de Reinforcement Learning (o simplemente RL). Pero, ¿Qué es el Reinforcement Learning?

2.0.2. Aprendizaje por Refuerzo (RL)

El **Reinforcement Learning** (RL), o Aprendizaje por Refuerzo, es una rama del aprendizaje automático no supervisado cuyo fundamento teórico se remonta a la **psicología conductista**, desarrollada en el siglo XX por autores como Thorndike (1911) y Pavlov (1927). En este ámbito, el aprendizaje se concibe como un proceso de refuerzo (las conductas seguidas de consecuencias positivas tienden a repetirse, mientras que las seguidas de consecuencias negativas tienden a extinguirse). Este principio fue adaptado al ámbito computacional con los trabajos de Sutton and Barto (1998), quienes formalizaron el marco de interacción agente–entorno bajo el mo-

[5] Existen reglas automáticas para seleccionar ℓ que minimizan un estimador del MSE de la varianza de largo plazo; ver Politis and White (2004).

delo de Proceso de Decisión de Markov (MDP). En finanzas, otros artículos como Moody et al. (1998) y Moody and Saffell (2001) aplicaron RL a estrategias de trading adaptativo, donde desmotraron su potencial para entornos no estacionarios y de alta volatilidad.

A diferencia del **aprendizaje supervisado**, donde se dispone de un conjunto de muestras (x_i, y_i) con las respuestas correctas previamente etiquetadas, en RL el agente no recibe instrucciones explícitas sobre la acción correcta en cada situación. En su lugar, debe descubrir una **política** de comportamiento que le permita maximizar la **recompensa acumulada** a partir de su propia experiencia e interacción con el entorno. A diferencia del aprendizaje supervisado, que depende de datos etiquetados previamente, el RL requiere únicamente una señal de recompensa, la cual puede ser escasa o diferida. En algunos casos, ambos enfoques se combinan en *aprendizaje por imitación*, donde una política inicial se entrena con datos supervisados y luego se refina mediante refuerzo. Esta diferencia hace que el RL sea especialmente adecuado para problemas de **toma de decisiones secuenciales**, donde las acciones presentes influyen en las oportunidades futuras.

Componentes fundamentales En RL, el proceso de aprendizaje se describe a través de la interacción entre:

- **Agente** ($\mathcal{A}$): la entidad que toma decisiones. En un contexto financiero, el agente representa el modelo de inversión que decide si comprar, mantener o vender activos.
- **Entorno** ($\mathcal{E}$): el sistema con el que el agente interactúa y que responde a sus acciones. En finanzas, el entorno está representado por el mercado y sus dinámicas de precios.
- **Estado** ($S_t \in \mathcal{S}$): la información que el agente observa en un momento dado, por ejemplo, un vector o matriz de información de mercado disponible en t, En esta tesis, S_t se construye como la **ventana de observación de retornos logarítmicos**:

$$S_t = O_t = \begin{bmatrix} R_{t-w+1}^{a_1} & \cdots & R_{t-w+1}^{a_N} \\ \vdots & \ddots & \vdots \\ R_t^{a_1} & \cdots & R_t^{a_N} \end{bmatrix},$$

donde para simplificar notación, $R_t^{a_i} = \ln\left(\frac{P_t^{a_i}}{P_{t-1}^{a_i}}\right)$ es el retorno logarítmico del activo i en el tiempo t, N es el número de activos y w la longitud de la ventana temporal.

- **Acción** ($A_t \in \mathcal{A}$): decisión tomada por el agente. En gestión de portafolios, esto puede ser vender, mantener o comprar un activo. En particular, en esta tesis, es un vector discreto con la decisión por activo (0 = vender, 1 = mantener, 2 = comprar).

- **Recompensa** ($g_t \in \mathbb{R}$): Es la cuantificación numérica de la calidad de la acción tomada. Por un lado, las recompensas positivas refuerzan comportamientos deseados aprendidos por el agente, mientras que recompensas negativas (castigos) desalientan comportamientos indeseados del agente. En mercados, puede definirse en función de métricas como el retorno, el ratio de Sharpe o el Expected Shortfall.

- **Política** ($\pi(a|s)$): Define la probabilidad de seleccionar una acción a dado un estado s. En la práctica, la política se implementa como una red neuronal parametrizada $\pi_\theta(a|s)$, la cual se actualiza de forma iterativa para maximizar la expectativa de la recompensa terminal:

$$\pi^* = \arg\max_{\pi_\theta} \mathbf{E}[R_T].$$ [6]

- **Factor de descuento** ($\gamma \in [0, 1]$): pondera la relevancia de recompensas futuras y pasadas.

Un aspecto esencial del RL es el balance entre la explotación y exploración. Donde esta se definen por:

- **Exploración**: búsqueda de nuevas combinaciones de decisiones de inversión que puedan ofrecer oportunidades no descubiertas.

- **Explotación**: reforzar aquellas decisiones que históricamente han generado un mejor desempeño ajustado por riesgo.

[6] Se dice que R_T depende de π_θ, aunque no directamente, sino como variable inducida a partir de las acciones y estados que se generan por la política π_θ

Durante el entrenamiento, este balance determina la eficiencia con que el agente aprende. Pues, una exploración insuficiente puede llevar a políticas subóptimas, mientras que una exploración excesiva ralentiza la convergencia. Los algoritmos modernos de RL (p. ej., PPO o DQN) incorporan este balance mediante técnicas como ε-greedy o entropía en la política estocástica.

El funcionamiento de un modelo de RL puede describirse de forma sencilla. Pues sigue un ciclo según los pasos:

1. El agente observa una ventana temporal $\{R^p_{-3}, R^p_{-2}, R^p_{-1}, R^p_0\}$ acotada hasta el presente (los retornos en este caso)
2. Toma una decisión basada en su política de comportamiento.
3. El entorno responde al agente con una recompensa o castigo.
4. El agente actualiza su política de comportamiento según la recompensa o castigo recibido.

Formalmente, el RL suele modelarse como un **Proceso de Decisión de Markov** (MDP), definido por la tupla:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma),$$

tal que satisface el **supuesto de Markov**, es decir:

$$\mathbf{P}(S_{t+1} \mid S_t, A_t, S_{t-1}, A_{t-1}, \dots) = \mathbf{P}(S_{t+1} \mid S_t, A_t).$$

En otras palabras, el estado futuro depende únicamente del estado y la acción actuales, y no de la historia completa.

El agente busca encontrar una política óptima π^* que maximice el **retorno esperado descontado**:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \quad \pi^* = \arg\max_{\pi} \mathbf{E}_\pi[G_t \mid S_t = s].$$

Para evaluar una política, se utilizan **funciones de valor**, las cuales cuantifican la calidad esperada de las decisiones bajo una política dada π:

$$V^\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s], \quad Q^\pi(s, a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a].$$

donde $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$ es el retorno descontado acumulado desde el tiempo t. En gestión de portafolios, $V^\pi(s)$ puede interpretarse como el valor esperado de mantener la actual composición del portafolio dado el estado de mercado s, mientras que $Q^\pi(s,a)$ representa el valor esperado de aplicar la acción a (por ejemplo, reequilibrar posiciones) en dicho estado.

Estas funciones se dicen que son de valor dado que miden la utilidad esperada de seguir una política π en términos de recompensas futuras descontadas. Estas ecuaciones, además satisfacen la ecuación de Bellman, que deriva directamente del principio de optimalidad de Bellman desarrollado en 1957 en Bellman (1957). Dicha ecuación, dice que todo valor esperado puede descompenerse en la recompensa inmediata sumada con el valor esperado esperado del siguiente estado. Formalmente,

$$V^\pi(s) = \sum_a \pi(a|s) \sum_{s'} \mathbf{P}(s'|s,a) \left[R(s,a) + \gamma V^\pi(s') \right],$$

y de manera análoga para la función de acción–valor:

$$Q^\pi(s,a) = \mathbb{E}_{s'} \left[R(s,a) + \gamma \sum_{a'} \pi(a'|s') Q^\pi(s',a') \right].$$

Estas ecuaciones son la base de los algoritmos de aprendizaje por refuerzo, pues permiten estimar iterativamente los valores esperados de las políticas y, en consecuencia, mejorar la toma de decisiones del agente.

En finanzas, el RL permite el modelamiento de la asignación de los pesos de los activos como un problema de decisiones en secuencia bajo incertidumbre. En específico, el agente recibe como estado datos de mercado, ejecuta acciones de sobre la asignación de pesos de los activos, y recibe recompensas basadas en métricas financieras de desempeño. A contraste de modelos estáticos (como el de Markowitz (1952) o Black and Litterman (1992)), el RL aprende continuamente y puede ajustar la política de asignación de activos en respuesta a cambios de régimen de mercado inclusive implícitos. Teniendo una ventaja mayor en entornos donde las distribuciones de retornos presentan los **"hechos estilizados"** presentados en el capítulo anterior.

A finales del siglo pasado, Moody et al. (1998) fueron pioneros en la aplicación del RL a la finanzas, e introdujeron el *Recurrent Reinforcement Learning* (RRL) en la optimización de métricas financieras como el ratio de Sharpe, marcando un hito en

la modelación financiera. Posteriormente, Moody and Saffell (2001) extendieron el enfoque a portafolios con varios activos, donde demostraron que el RL posee capacidad de adaptación en entornos no estacionarios. Más recientemente, Fischer (2018) y Bai et al. (2024), destacan el uso de algoritmos modernos (PPO, DDPG, SAC) para la gestión dinámica de portafolios, market making y cobertura mediante derivados financieros, con la integración de arquitecturas profundas (CNN, LSTM) y mecanismos de exploración estocástica.

Un problema existente en la utilización del RL a las finanzas, es que el MDP asume *estacionariedad* y *observabilidad* del estado. Sin embargo, los mercados reales son no estacionarios y parcialmente observables. Una posible solución propuesta por Moody and Saffell (2001) es la de la utilización de políticas estocásticas (p.ej., entropía en PPO) para la adaptación a cambios de régimen.

El entorno financiero se puede representar como un MDP, donde el agente observa estados de mercado, ejecuta acciones de inversión y recibe recompensas asociadas al desempeño del portafolio. Sin embargo, en la práctica los mercados son no estacionarios y parcialmente observables, lo que implica que las dependencias entre activos no siempre son explícitamente conocidas ni estables en el tiempo. Para mitigar lo anterior, siguiendo el planteamiento de Moody and Saffell (2001), se introduce la **hipótesis de independencia condicional por activo** definida a continuación:

Definición 11 Sea $\mathbf{R}_t = (R_{1,t}, R_{2,t}, \ldots, R_{N,t})$ el vector de retornos de N inversiones en el periodo t, y sea s_t el estado del mercado observado. Se dice que los retornos cumplen la hipótesis de independencia condicional por inversión si:

$$\mathbf{P}(\mathbf{R}_t \mid s_t) = \prod_{i=1}^{N} \mathbf{P}(R_{i,t} \mid s_t).$$

Esta condición implica que, dados los estados de mercado s_t, los retornos de las inversiones son condicionalmente independientes entre sí.

Definición 12 En un MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, una **política** es una función:

$$\pi : \mathcal{S} \to \Delta(\mathcal{A}),$$

donde $\Delta(\mathcal{A})$ representa el conjunto de distribuciones de probabilidad sobre el espacio de acciones $\mathcal{A}$. Así, $\pi(a|s)$ denota la probabilidad de seleccionar la acción a en el estado s. La política óptima π^* maximiza el retorno esperado descontado:

$$\pi^* = \arg\max_{\pi} \ \mathbf{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \right].$$

En el entorno sobre el cual se trabaja en esta tesis, la política estocástica $\pi_\theta(a \mid O_t)$ puede factorizarse como

$$\pi_\theta(a \mid O_t) \;=\; \prod_{i=1}^{n} \pi_\theta^{(i)}(a_i \mid O_t),$$

bajo la **hipótesis de independencia condicional por activo**, que asume que los retornos son condicionalmente independientes dados los estados de mercado observados. Esta simplificación, se utiliza frecuentemente en modelos de RL para portafolios grandes, en el mismo artículo de Moody and Saffell (2001) utilizan dicha simplificación. Su justificación, para el contexto de esta tesis, proviene de que:

- El espacio de acciones conjuntas $\mathcal{A} = \{0, 1, 2\}^n$ crece de forma exponencial con n, lo que hace inviable modelar dependencias completas en problemas realistas. La factorización reduce la complejidad de $\mathcal{O}(3^n)$ a $\mathcal{O}(n)$, permitiendo poder extender este modelo a más número de activos en cartera, tal como lo propone Kolm et al. (2020) en su artículo. La factorización ha motivado enfoques alternativos basados en espacios de acción continuos, como el propuesto por Li (2019), donde las decisiones se modelan directamente como pesos proporcionales de inversión.

- En entornos con limitancia de datos y dinámicas no estacionarias, asumir independencia condicional actúa como una forma de regularización implícita, reduciendo la complejidad del modelo y evitando que aprenda patrones específicos de la muestra, conllevando a un mayor riesgo de sobreajuste. Este tipo de sesgos, son comunes en deep learning y, según (Goodfellow et al., 2016), contribuyen a mejorar la estabilidad del entrenamiento en presencia de ruido y alta dimensionalidad.

- Moody and Saffell (2001) muestran que las políticas factorizadas pueden capturar correlaciones indirectamente a través del del estado O_t, sin la necesidad de modelar de forma explícita todas las interdependencias que existan.

Aunque la factorización omita dependencias cruzadas explícitas, estas pueden incorporarse indirectamente en la representación de estado. Para entornos donde dichas dependencias son críticas, Buehler et al. (2019) ha propuesto extensiones con enfoques de políticas conjuntas mediante arquitecturas profundas, tales como Deep Hedging para la cobertura de derivados financieros.

Matemáticamente, una política determinista es una función $\mu : \mathcal{S} \to \mathcal{A}$, mientras que una política estocástica define una familia de distribuciones categóricas sobre $\{0, 1, 2\}^n$. En cada paso t, el agente puede explorar muestreando una acción $A_t \sim \pi_\theta(\cdot \mid O_t)$, o explotar eligiendo la acción más probable según su política $A_t = \arg\max_a \pi_\theta(a \mid O_t)$, recibiendo así una recompensa g_t (el índice de Sharpe en este caso) que refuerza el comportamiento del agente de RL, según:

- Vender si el activo cae (recompensa positiva por evitar pérdidas).
- Comprar si sube (beneficio por aprovechar la subida del precio).
- Mantener si el movimiento esperado es marginal respecto a un umbral de decisión τ y a los costes de transacción.

Capítulo 3

Adquisición de los Datos

El objetivo de este capítulo es describir el proceso de adquisición y preparación de los datos que servirán como insumo para el modelo de RL. Aunque la metodología es aplicable a un amplio conjunto de activos transados en los mercados de capitales, se ilustra inicialmente con el caso de las criptomonedas, ETF y algunas acciones, especificadas a continuación.

Los precios se obtienen desde *Yahoo Finance* para los tickers **BTC-USD**, **ETH-USD**, **SPY**, **TLT**, **AAPL** y **JPM**, en frecuencia diaria, ajustados por dividendos y *splits*. La ventana de análisis es **junio de 2018 a agosto de 2025** se selecciona por tres razones principales:

1. **Cobertura de ciclos completos**: abarca al menos dos ciclos de mercado bien diferenciados en criptoactivos y renta variable (caídas 2018, shock COVID-19 en 2020, rally 2020–2021, corrección 2022 y recuperación 2023–2025), lo que permite entrenar y evaluar bajo regímenes heterogéneos.

2. **Eventos de alta volatilidad y riesgo de cola**: la ventana temporal incluye episodios extremos relevantes para métricas como VaR y ES, condición necesaria para testar políticas de *Reinforcement Learning* en contextos con colas pesadas.

3. **Balance entre extensión y estabilidad**: siete años de datos diarios entregan una cantidad de datos suficiente para particionar en *entrenamiento/validación/testeo* sin fuga de información. Pues, como se verá más adelante, cada partición contiene al menos un año de información.

En conjunto, esta ventana mejora la evaluación fuera de muestra al incorporar distintas fases económicas, y mejora la generalización del agente al cubrir múltiples regímenes y shocks (como la pandemia en contraste con la situación actual), condición clave cuando el objetivo es optimizar medidas de riesgo de cola como el Expected Shortfall.

Además, se tienen en cuenta los siguientes puntos para la estructuración del dataset a utilizar como fuente de información:

- **Zona horaria:** se homogeniza a UTC y se alinean días hábiles por intersección de calendarios; feriados sin precio se completan por *forward-fill*. Dicha normalización permite evitar asincronías en precios entre mercados, y mantiene comparabilidad entre retornos, siendo esencial para la estimación, tal como lo postula (Lo and MacKinlay, 1990).
- **Ajustes corporativos:** se usan precios *adjusted close* para acciones y ETF; en cripto, se emplea *close* simplemente, dado que las criptomonedas no poseen dividendos ni splits.
- **Datos faltantes:** Se descartan tramos con lagunas superiores a tres días hábiles consecutivos, dado que tales discontinuidades pueden distorsionar la estimación de retornos y volatilidades (Tsay, 2010).[1]
- **Versionado:** se fijan semillas aleatorias mediante (`random_state`).

3.1. Criptomonedas

Se seleccionaron las criptomonedas **Bitcoin (BTC)** y **Ethereum (ETH)** debido a:

1. Su alta liquidez, que reduce el sesgo por iliquidez
2. La amplia disponibilidad de datos en diferentes frecuencias (diaria y horaria)

[1] El RL puede lidiar con esto dado que aprende a decidir bajo observabilidad parcial. Si no hay información, el agente ajusta la política según lo que sí observa. Al rellenar los datos, se introduciriían valores artificiales que podrían distorsionar los retornos reales.

3. Su marcada volatilidad, que permite evaluar el desempeño de modelos en contextos de riesgo no normal.

La construcción inicial de una cartera de inversión, o simplemente portafolio de activos, se realizó utilizando 2 activos financieros, dados el par de criptomonedas mencionadas y otros pares activos. Se descargó el historial de precios en frecuencia diaria, cubriendo el periodo mencionado anteriormente. Se emplean retornos logarítmicos diarios sincronizados.Dado que el mercado de criptomonedas opera las 24 horas del día, todos los días de la semana, mientras que las acciones y ETF se rigen por el calendario bursátil, la sincronización de las series se efectúa considerando únicamente la intersección de sus periodos temporales comunes. En su artículo, Cushing (2000) evidencia que los retornos de cierre a cierre reducen la influencia del ruido de microestructura intradía (como el bid-ask bounce, que son pequeños saltos falsos en el precio por la diferencia entre compra y venta) y distorciones asociadas al rebalance de fin de jornada, facilitando la comparación entre clases de activos que tengan distinta liquidez o frecuencias de sus cotizaciones en el mercado.

Posteriormente, se calculan los retornos logarítmicos de cada uno de los activos que componen el portafolio, para así obtener la correlación entre sus retornos. En este caso, la correlación entre Bitcoin y Ethereum es elevada ($\rho \approx 0.83$), lo que refleja que ambas criptomonedas comparten gran parte de sus dinámicas de riesgo y retorno. Esto tiene implicancias en la diversificación del portafolio, ya que una alta correlación reduce los beneficios de dispersión del riesgo. Considerando que dichos retornos serán utilizados para el entrenamiento de un modelo de reinforcement learning, es importante visualizar la serie, retornos y los valores atípicos de los datos. En tal caso, se tiene que:

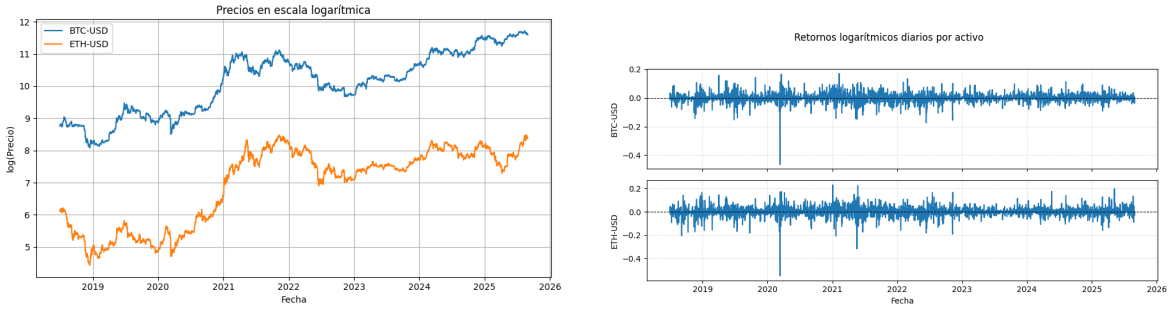


Figura 3.1. Serie de precios (izquierda) y retornos logarítmicos (derecha) de Bitcoin y Ethereum.

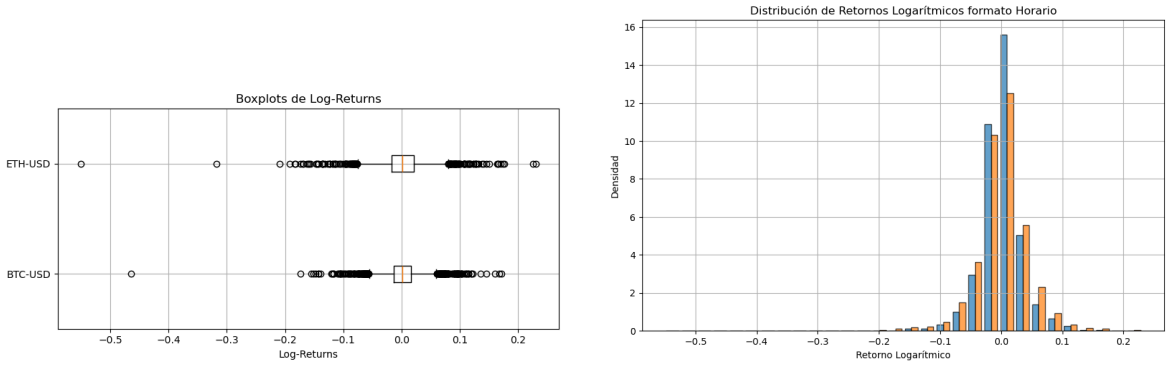


Figura 3.2. Boxplots (izquierda) e histogramas de distribución (derecha) de retornos de Bitcoin y Ethereum.

En las Figuras 3.1 y 3.2, se observa la presencia de numerosos valores atípicos (*outliers*, por ejemplo aquel -0,5 para ETH y el -0,46 para BTC en Marzo de 2020 o el 0,31 de ETH de mayo de 2021), los cuales son movimientos de precios extremos, que en finanzas representan eventos de alta volatilidad, tales como caídas abruptas o subidas pronunciadas.
La identificación de valores atípicos es relevante porque:

- Sugiere la presencia de distribuciones con colas pesadas
- Impactan directamente en métricas de riesgo como el *Value-at-Risk (VaR)* y el *Expected Shortfall (ES)*
- Determinan la magnitud de recompensas extremas que un agente de *Reinforcement Learning* puede recibir, afectando su proceso de entrenamiento.

Activo	Media	Desv.	Asimetría	Curtosis	VaR_{95}	ES_{95}	VaR_{99}	ES_{99}
BTC–USD	0.0011	0.0334	-1.03	17.36	-0.051	-0.079	-0.092	-0.132
ETH–USD	0.0009	0.0442	-0.94	12.84	-0.068	-0.107	-0.134	-0.181

Tabla 3.1. Estadísticos descriptivos y medidas de riesgo empírico para criptomonedas.

En la Tabla anterior, la columna *Curtosis* corresponde al exceso de curtosis, definido como:

$$\text{Curtosis} = \frac{\mathbf{E}[(X-\mu)^4]}{\left(\mathbf{E}[(X-\mu)^2]\right)^2} - 3.$$

Dada la presencia de bastantes valores atípicos, sugiere que su distribución sería una no normal. Lo anterior se observa en la Figura 3.2 y en la tabla 3.1.

Como se observa en la Tabla 3.1, la curtosis de exceso de ambas criptomonedas es considerablemente superior a cero, alcanzando valores de 17.36 para Bitcoin y 12.84 para Ethereum. Indicando una distribución leptocúrtica, caracterizada por una alta concentración de observaciones en torno a la media y colas más pesadas que las de una distribución normal. Sugiriendo una mayor probabilidad de ocurrencia de eventos extremos, coherente con los picos observados en la Figura 3.2. Asimismo, la asimetría negativa para ambos activos indica una leve desviación de las distribuciones hacia las pérdidas pronunciadas, reforzando la presencia de riesgo de cola en el lado izquierdo de las distribuciones de retornos.

Esta característica, ampliamente documentada en la literatura financiera (por ejemplo, Mandelbrot (1963); Fama (1965); Cont (2001)), justifica el uso de distribuciones alternativas como la t–Student, distribuciones asimétricas o modelos de teoría de valores extremos, como la Distribución Generalizada de Pareto (GPD), para capturar el riesgo de cola.

A partir del contraste de normalidad Shapiro Wilk, se rechazó la hipótesis de normalidad en ambos activos (*BTC–USD*: $W = 0.90$, $p < 0.001$; *ETH–USD*: $W = 0.91$, $p < 0.001$), posiblemente debido a la presencia de colas pesadas o asimetrías como las visibles en la Tabla 3.1. En consecuencia, una distribución t–Student logra un ajuste más adecuado (con grados de libertad estimados numéricamente para cada activo mediante el método de máxima verosimilitud $\nu_{BTC} \approx 2.42$ y $\nu_{ETH} \approx 2.78$), lo cual es consistente con lo dicho en el capítulo anterior respecto a colas más pesadas.[2] No obstante, dado que los retornos financieros suelen exhibir asimetrías y colas extremas, también podría considerarse el ajuste mediante distribuciones *skewed* o técnicas de *Extreme Value Theory* (EVT) para modelar adecuadamente los riesgos de cola. Esto es particularmente relevante para métricas como el Expected Shortfall.[3]

[2] Las comparaciones de verosimilitud muestran mejoras en el ajuste (Δ AIC ≈ -700 para ambos)

[3] Si bien la EVT permite modelar eventos extremos de manera más específica, excede el alcance de esta tesis, cuyo objetivo principal es caracterizar el comportamiento general de los retornos y su impacto en la toma de decisiones del agente

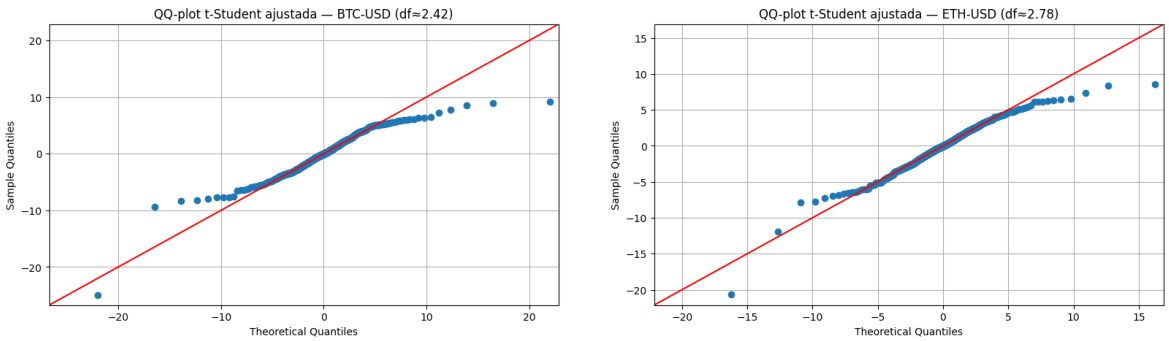


Figura 3.3. QQ Plots de Bitcoin y Ethereum para una distribución t-student

La Figura 3.3 muestra que los cuantiles empíricos se acercan más a los cuantiles teóricos de la distribución t-Student ajustada, confirmando que los retornos presentan colas más pesadas. Esto es consistente con los resultados de pruebas estadísticas de normalidad, donde se rechazó la hipótesis nula de normalidad.

3.2. ETF

Para capturar factores sistemáticos de renta variable y riesgo de tasas de interés de largo plazo, se incorporan dos *Exchange-Traded Funds*, que son instrumentos financieros líquidos y representativos, agrupando varios activos (acciones, bonos, materias primas, etc.) y se negocian en la bolsa de valores, al igual que las acciones. Los utilizados para esta tesis, son:

- **SPDR S&P 500 (SPY)**: proxy del mercado accionario estadounidense de gran capitalización (cercano a la *beta* de mercado [4]), alta liquidez y bajo costo relativo de transacción. Este ETF, permite evaluar el agente en un activo *benchmark* con microestructura estable.
- **iShares 20+ Year Treasury Bond (TLT)**: exposición a bonos del Tesoro de EE. UU. de *larga duración*, altamente sensible a variaciones en la pendiente y nivel de la curva de tasas. El cual aporta un comportamiento defensivo en shocks al ser un instrumento de renta fija, permitiendo un mayor grado de diversificación frente a renta variable y criptoactivos.

[4]La beta es una métrica financiera que mide la sensibilidad de un activo respecto al mercado

Una propiedad empírica relevante para la construcción del portafolio es la **correlación negativa** entre el proxy de mercado accionario estadounidense (SPY) y la exposición a bonos del Tesoro de EE. UU. de larga duración (TLT). En la ventana temporal usada, la correlación de los retornos logarítmicos diarios es $\rho \approx -0,15$. El signo negativo refleja el *trade-off* o compensación estructural entre *beta* de mercado (crecimiento/acciones) y **duración larga** (sensibilidad a tasas). Este patrón es coherente con episodios de *flight-to-quality*, fenómeno financiero en que los inversionistas venden activos de alto riesgo (como la renta variable) y se refugian en activos de renta fija de bajo riesgo (como los bonos del tesoro de Estados Unidos) deibo a la incertidumbre, en los que caídas en renta variable coinciden con alzas de precios en bonos de larga duración. Aunque la magnitud es *moderada* en promedio, esta relación es variante en el tiempo (o como se dice en finanzas, *time varying*), pues se intensifica en shocks macro de tasas e inflación y puede atenuarse en fases de apetito por riesgo. En nuestro contexto, esta dependencia negativa:

- **Mejora la diversificación** dado que introduce cierta compensación entre activos de riesgo de renta variable (acciones en este caso) y activos defensivos de renta fija (bonos).
- **Reduce el riesgo de cola conjunto** (y, por ende, el *Expected Shortfall* de la cartera) gracias a la diversificación cuando las coberturas de duración actúan de forma efectiva.
- Proporciona un **entorno heterogéneo** para el agente de *Reinforcement Learning*, que debe aprender políticas robustas frente a cambios de régimen y correlaciones que no son constantes en el tiempo.

La combinación SPY–TLT introduce un **trade-off o compensación de crecimiento vs. duración** y correlaciones *time-varying o variantes en el tiempo* que permiten testear si las políticas aprendidas se adaptan a cambios de régimen (inflación/tasas altas vs. relajación monetaria).

Dada la correlación negativa en sus retornos, es esperable que un portafolio compuesto con ambos activos tenga un riesgo de cola menor, por lo que una política basada en ES, podría eventualmente tener un comportamiento más errático. Pues, por naturaleza, disminuye el riesgo de caídas extremas del portafolio. Sumado a lo

anterior, si se observa la Figura 3.5, se podrá observar que este par de activos compone (de los 3) aquellos con menor cantidad y magnitud de valores atípicos.

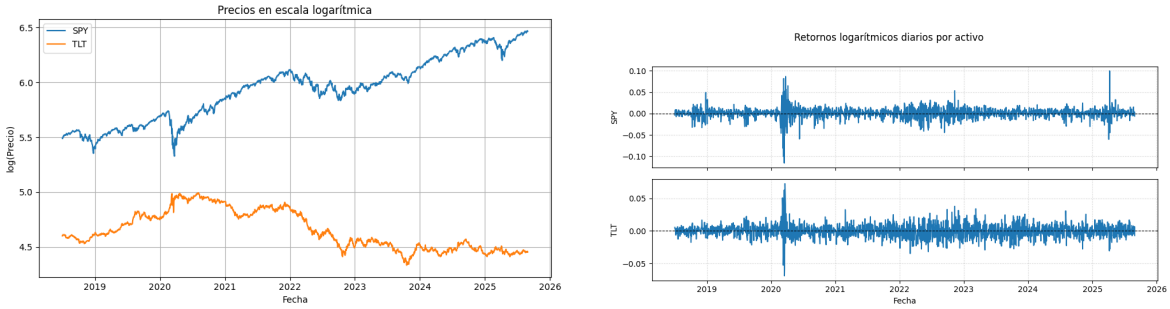


Figura 3.4. Serie de precios (izquierda) y retornos logarítmicos (derecha) de SPY y TLT.

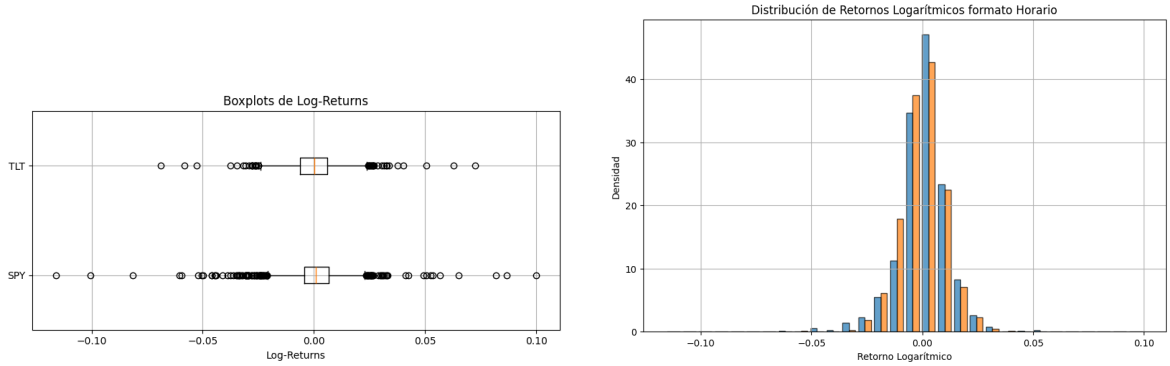


Figura 3.5. Boxplots (izquierda) e histogramas de distribución (derecha) de retornos de SPY y TLT.

En la Figura 3.4 se nota que SPY mantiene una tendencia alcista, mientras que el TLT tiene un comportamiento más plano y descendente desde 2020, teniendo un comportamiento inverso con el otro ETF lo cual es coherente con su correlación negativa. Los retornos de la Figura 3.4 muestran aumentos de volatilidad en distintos periodos, evidenciando una posible heterocedasticidad que sería coherente con lo presentado en el capítulo anterior respecto a volatilidad de activos.

Las observaciones anteriores son posible de inferir a partir del histograma de la Figura 3.5, que son consistentes con las estadísticas de la Tabla 3.2, pues la escala del gráfico es menor y no se observa gran densidad en retornos mayores a 0.05 en magnitud, siendo también menos asimétrica visualmente en comparación con las criptomonedas, pero aún siendo asimétrica. Pues ambos activos muestran una concentración alta de retornos en torno a cero y colas más pesadas que las de una distribución

normal, lo cual se refleja en los valores positivos de curtosis excesiva (13.47 para SPY y 4.48 para TLT).

Activo	Media	Desv.	Asimetría	Curtosis	VaR_{95}	ES_{95}	VaR_{99}	ES_{99}
SPY	0.0005	0.0126	-0.55	13.47	-0.0186	-0.0309	-0.0348	-0.0547
TLT	-0.0001	0.0103	0.06	4.48	-0.0165	-0.0224	-0.0249	-0.0340

Tabla 3.2. Estadísticos descriptivos y medidas de riesgo empírico para ETFs.

Para finalizar, y confirmar las observaciones realizadas para las criptomonedas, el test de Shapiro-Wilk rechazó la hipótesis de normalidad para los retornos de ambos activos, teniendo SPY una desviación más pronunciada ($W = 0,87$ y $p < 0,001$) mienrtas qur TLT muestra menos discrepancia ($W = 0,97$ y $p < 0.001$). Aún así, la distribución t-Student se mostró con una mejora significativa respecto a la normal, con grados de libertad de 2,77 para SPY y de 6,08 para TLT, concluyendo la presencia de colas más pesadas y un buen ajuste de ellas, como se aprecia en la Figura 3.6.[5]

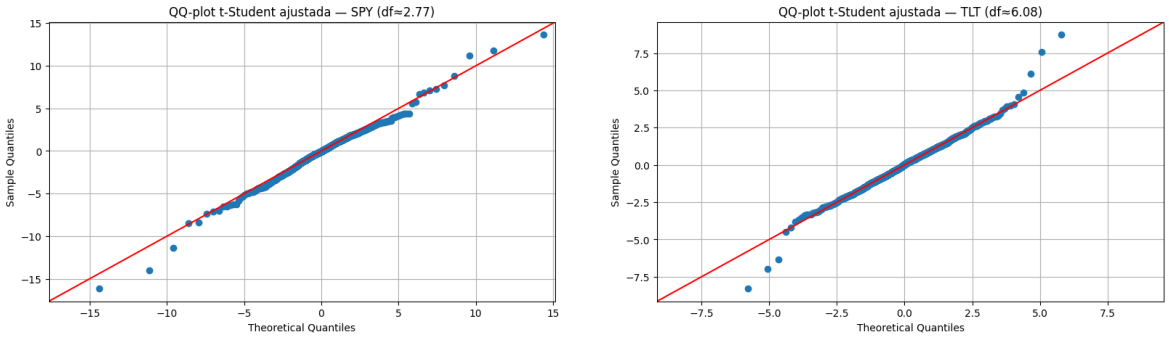


Figura 3.6. QQ Plots de SPY y TLT para una distribución t-student

3.3. Acciones

Para incorporar riesgo idiosincrático (de un activo en particular) y sesgos sectoriales, se consideran dos empresas de gran capitalización estadounidenses con perfiles diferenciados:

- **Apple Inc. (AAPL)**: Una empresa tecnológica con alto componente de crecimiento e intensidad en intangibles. Su volatilidad y sensibilidad a expectati-

[5]Para los ETF, también se produce de que Δ AIC < 0

vas de tasas/innovación complementan el sesgo tecnológico capturado parcialmente por el ETF SPY mencionado en la sección anterior.

- **JPMorgan Chase & Co. (JPM)**: El cual es un líder global de servicios financieros, con una gran corporación bancaria en Estados Unidos, es una acción representativa del sector financiero. Por lo que, su desempeño depende de la actividad crediticia, márgenes por tasas y regulación. Introduce exposición a ciclo financiero real y a la curva de tipos.

AAPL y JPM aportan **heterogeneidad sectorial** (tecnología vs. banca), aportan una nueva arista la matriz de correlaciones de ETF y cripto, y permiten evaluar la **transferencia de políticas** del agente entre activos con distintas dinámicas de volatilidad, eventos corporativos y sensibilidad macro. De hecho, su heterogeneidad se puede ver a partir de la baja correlación existente entre ambos activos. Pues en este caso la correlación entre sus retornos logarítmicos diarios es moderada y positiva $\rho \approx 0,43$, indicando que existe cierta relación entre sus movimientos, aunque no lo suficientemente fuerte como para indicar dependencia estructural. Desde un punto de vista estadístico, el valor de las correlaciones entre los activos y ETF, podría considerarse bajo. Sin embargo, desde el punto de vista económico refleja que ambos activos responden parcialmente a factores comunes del ciclo financiero[6], manteniendo al mismo tiempo comportamientos propios de cada sector, permitiéndole así no tener una correlación tan alta. Sus retornos son menos volátiles en comparación a las criptomonedas, pues en su boxplot de la Figura 3.8, se observa que los valores atípicos mantienen una escala menor en comparación con los de las criptomonedas. En la misma Figura 3.8, se puede notar que sus retornos logarítmicos, muestran colas menos pesadas en comparativa a las criptomonedas.

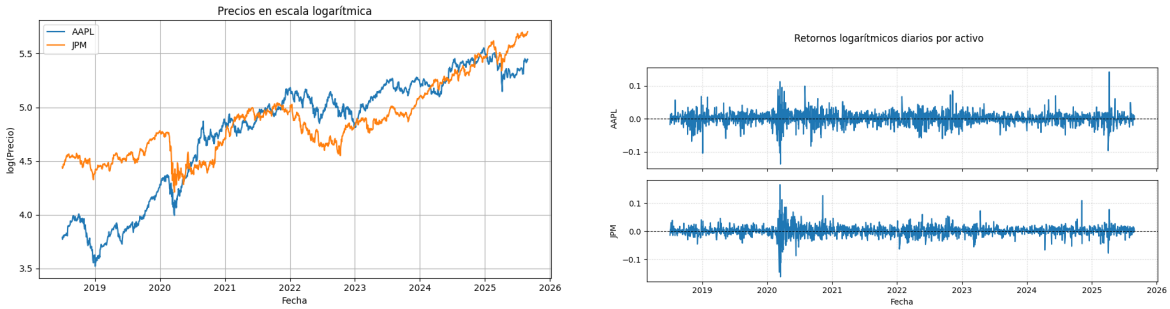


Figura 3.7. Serie de precios (izquierda) y retornos logarítmicos (derecha) de AAPL y JPM.

[6]Por ejemplo, cambios en la tasa de política monetaria, apetitos por riesgo, crisis globales

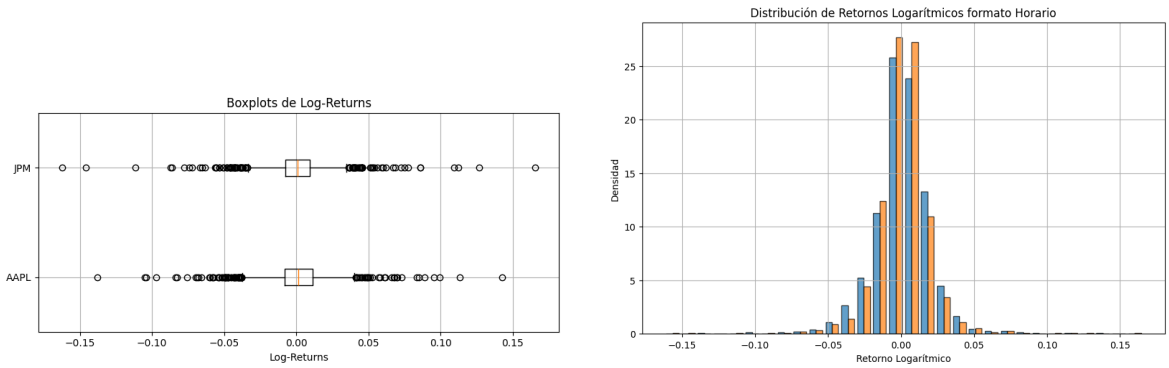


Figura 3.8. Boxplots (izquierda) e histogramas de distribución (derecha) de retornos de AAPL y JPM.

En la Figura 3.7 se observa una tendencia alcista para ambos activos, con un leve quiebre en el periodo de tensión dado por la pandemia, el cual se vió más marcado para JPM por su lenta recuperación en comparación a AAPL. También, en la Figura 3.7 se observa la existencia de picos de volatilidad en esos periodos, sin ser perfectamente alineados entre el máximo drawdown del periodo (caída más baja) y la máxima subida. El comportamiento del precio entre ambas series es relativamente similar, ambos activos resienten los shocks económicos más fuertes, aunque persiste una componente idiosincrática sectorial, lo que es coherente con su correlación moderada.

Para finalizar el análisis de las acciones escogidas, los contrastes de normalidad de Shapiro–Wilk rechazaron la hipótesis de normalidad en ambos activos, con AAPL mostrando una desviación moderada ($W = 0.93$, $p < 0.001$) y JPM una más pronunciada ($W = 0.88$, $p < 0.001$). En ambos casos, la distribución t–Student obtuvo un mejor ajuste respecto a la normal, con grados de libertad de 3.34 para AAPL y 2.94 para JPM, indicando exactamente lo mismo que para los otros pares de activos (colas más pesadas y una mayor propensión a eventos extremos, especialmente en el sector bancario), en la Figura 3.9 se observa que la distribución t-Student ofrece un mejor ajuste para ambos activos.[7]

[7]De igual forma, se verifica que $\Delta\text{AIC} < 0$, confirmando la mejor adecuación del modelo t–Student.

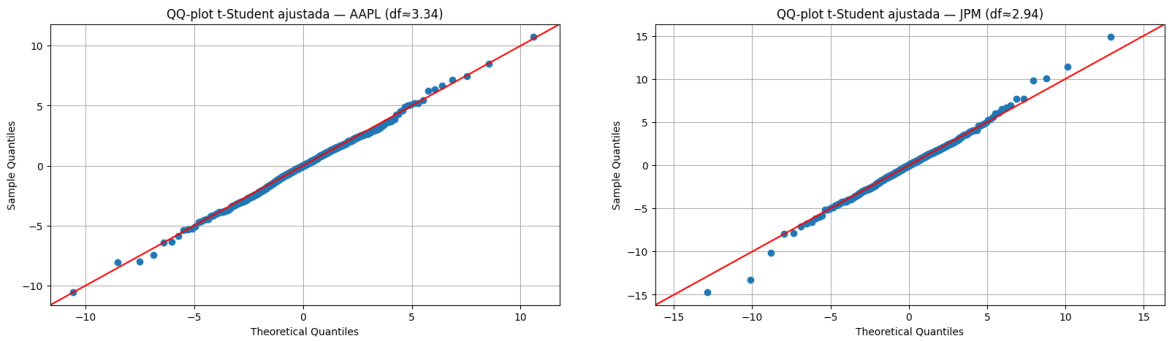


Figura 3.9. QQ Plots de AAPL y JPM para una distribución t-student

Los diagnósticos realizados durante la exploración de los datos, confirmaron propiedades clave para el modelado del riesgo de cola. Por ejemplo, la presencia de **leptocurtosis**, **outliers** frecuentes y **agrupamiento de volatilidad** en cripto. También, para el caso de los ETF, se mostró la existencia de **correlaciones negativas** entre SPY y TLT, y **heterogeneidad sectorial** entre AAPL y JPM.

En términos de implicancias para el resto de la tesis, se tiene que:

- La presencia de colas pesadas y asimetrías justifican el uso de métricas coherentes al riesgo de cola como el **Expected Shortfall (ES)** y motiva el uso de estimadores de volatilidad/ES *ex-ante* consistentes con distribuciones no normales. Para lograr utilizar el riesgo o volatilidad como una variable también.
- La **correlación negativa SPY–TLT** sugiere que la diversificación efectiva depende del régimen de mercado. Pues a pesar de sectores muy distintos, la magnitud de la correlación no era alta.

Sin embargo, con este enfoque se presentan algunas limitaciones, como por ejemplo que:

- El uso de precios *close-to-close*[8] no captura microestructura intradía. Pues ya se vió que el mercado de criptomonedas permanece en apertura durante todo las 24 horas.
- Gracias al descarte de tramos con lagunas superiores a tres días hábiles, existen posibles sesgos de supervivencia frente a empresas que derivaron en quiebra por liquidez.

[8]Utilización de precios de cierre ajustados, para mercados que no operan las 24 horas.

Capítulo 4

Metodología

El objetivo de esta tesis es generar un modelo de Reinforcement Learning que aprenda políticas de inversión sobre un portafolio de activos financieros, realizando cambios en la asignación de activos dentro de un portafolio, mediante laoptimización de métricas de desempeño financiero ajustadas por riesgo. En particular, la metodología implementada en este trabajo utiliza como criterio de recompensa el **índice de Sharpe**, presentado en el capítulo 2. De este modo, el agente busca aprender estrategias que no sólo maximicen la rentabilidad, sino que lo hagan en términos relativos al riesgo asumido, lo cual resulta especialmente relevante en mercados con alta volatilidad. En ese enfoque, en Vittori et al. (2020) realizan un estudio de la aplicación del RL a la gestión de derivados financieros, incorporando métricas de riesgo a la función de recompensa del modelo, obteniendo que dicha incoporación permite que el agente adapte la política de inversión frente a condiciones cambiantes, buscando priorizar la estabilidad del portafolio y la preservando el capital invertido.

Si bien en la práctica los modelos de trading enfrentan costos de transacción y fricciones de mercado, al momento de la escritura de esta tesis han empezado a masificarse plataformas que ofrecen transar sin costo en mercados de capitales. Por lo tanto, la comisión por transacción funciona como un hiperparámetro, ajustable según la plataforma que se utilice para trading.

4.1. Diseño del Entorno de RL

El enfoque se centra en evaluar la capacidad del agente de RL de generar retornos ajustados por riesgo. Para implementar este modelo, es necesario construir un entorno virtual sobre el cual el agente interactúe. Dicho entorno provee las reglas del sistema, como por ejemplo: la información recibe el agente, las acciones puede tomar, cómo se acumulan los resultados y qué señal de recompensa recibe.

El entorno se compone de:

- **Observaciones**: ventanas deslizantes de retornos logarítmicos de los activos, aplanadas en un vector de dimensión $w \cdot N$, donde w es el tamaño de la ventana y N el número de activos.
- **Acciones**: representan direcciones de exposición o decisiones discretas por activo, con tres posibilidades: -1 (posición corta o venta), 0 (posición neutra o mantener) o $+1$ (posición larga o comprar).
- **Recompensa**: al finalizar cada episodio se calcula la recompensa mediante el índice de Sharpe anualizado de los retornos acumulados del portafolio durante el episodio.
- **Funciones auxiliares**: tales como el reseteo del entorno, la gestión del horizonte máximo de pasos y la selección del índice inicial del episodio.

A continuación, se desglosará con mayor detalle el funcionamiento de cada uno de estos componentes.

4.1.1. Espacio de Observación y Decisión

Como se mencionó previamente y como es común en los modelos de predicción en finanzas, siguiendo la estructura recurrente de Fischer and Krauss (2018), el modelo de RL necesita datos observables para tomar decisiones, este espacio de observación representa un subconjunto del total de información disponible. Los datos vendrán dados por los retornos logarítmicos de activos dentro del portafolio (en este caso,

dos), los cuales forman una matriz de observaciones de tamaño $w \times N$, donde N representa el número de activos y w el número de observaciones de retornos que se tienen por activo dentro del portafolio.

Para capturar la dependencia temporal de los retornos, es necesario entregarle al agente un conjunto de observaciones (retornos logarítmicos) que actúen como historial (llamado *window* y denotada por O_t). Esta *ventana* es de la forma $w \times N$ y el agente la observa en cada episodio y la utiliza como fuente de información para tomar una decisión. Matemáticamente, O_t es una matriz de la forma:

$$O_t = \begin{bmatrix} R_{t-w+1}^{a_1} & R_{t-w+1}^{a_2} & \cdots & R_{t-w+1}^{a_N} \\ \vdots & \vdots & \ddots & \vdots \\ R_t^{a_1} & R_t^{a_2} & \cdots & R_t^{a_N} \end{bmatrix},$$

donde $R_s^{(i)} = \ln\left(\frac{P_s^{(i)}}{P_{s-1}^{(i)}}\right)$ corresponde al retorno logarítmico del activo i en el periodo s.

Si $t < w$ (es decir, durante el inicio de cada episodio), los datos faltantes se rellenan con ceros:

$$O_t(i,j) = 0 \quad \text{para} \quad i \leq w - t,\ j = 1, \ldots, N.$$

Donde lo anterior, podría ocurrir por ejemplo en caso en donde el modelo tome una ventana al inicio de los datos (no pudiendo tomar una ventana del tamaño w por la limitación de datos). Esta es una técnica ampliamente utilizada en el modelamiento de series financieras con aprendizaje por refuerzo, pues se utilizan datos secuenciales. En Jiang et al. (2017) se explica que dicha técnica permite evitar extrapolaciones que perturben el entrenamiento del modelo al inicio.

En este punto es necesario transformar la ventana $O_t \in \mathbb{R}^{w \times N}$ desde una matriz de dimensión $w \times N$ en un vector unidimensional de largo $w \cdot N$. Este paso, denominado *aplanamiento* (o *flattening* según la literatura de RL), se define como

$$\text{obs}_t = \text{vec}(O_t) \in \mathbb{R}^{w \cdot N}.$$

En la Figura 4.1 se muestra, a la izquierda, la ventana O_t como una matriz de retornos de tamaño $w \times N$, y a la derecha, se muestra cómo esa matriz se aplana en

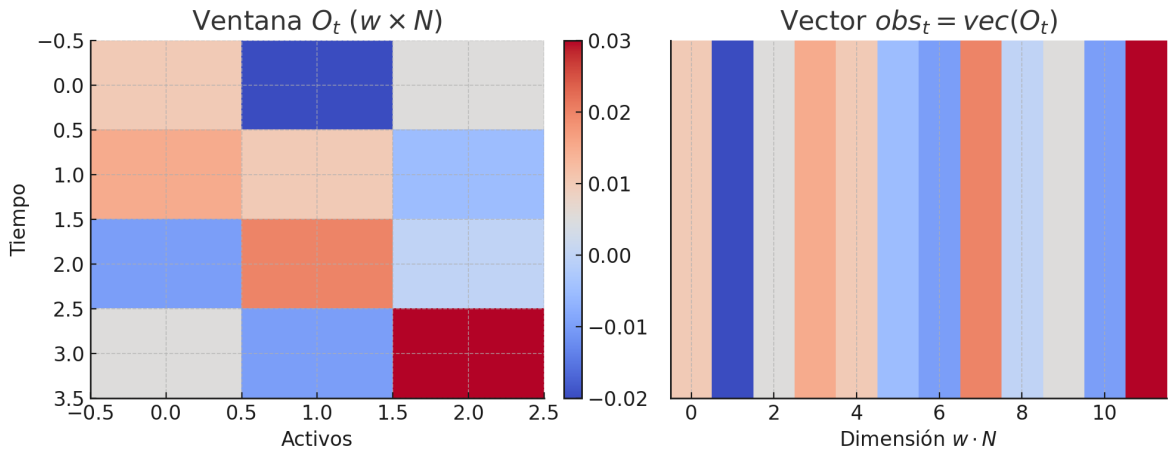


Figura 4.1. Aplanamiento de Obs

un vector de dimensión $w \cdot N$, que es la observación que recibe el agente.

Las principales razones por las cuales se aplica este procedimiento es que, en primer lugar, desde el punto de vista computacional, la mayoría de las bibliotecas de *Reinforcement Learning* (como `Gymnasium` Brockman et al. (2016) y `Stable-Baselines3` Raffin et al. (2021)) y de aprendizaje profundo están diseñadas para trabajar con vectores de entrada de dimensión fija, pues las políticas paramétricas empleadas en estos entornos son redes densas, que esperan vectores de tamaño constante. Representar el estado como un vector facilita la compatibilidad con capas densas (*fully connected*), la operación en *batch* y la propagación hacia atrás en el entrenamiento. En segundo lugar, desde el punto de vista metodológico, el aplanamiento permite condensar la información temporal y cross-asset de la ventana en una única representación de tamaño constante, independientemente de la longitud de la serie histórica.

Esto asegura que el espacio de observación quede formalmente definido como

$$\mathcal{O} = \left\{ \text{obs} \in \mathbb{R}^{w \cdot N} \right\},$$

lo que mantiene fijo el tamaño de la entrada al agente incluso cuando $t < k$ (en cuyo caso se aplica relleno con ceros).

Este tipo de representación es un estándar en la literatura de aprendizaje por refuerzo. Por ejemplo, Mnih et al. (2015) introducen el uso de vectores y tensores fijos como entradas al agente en el marco de *Deep Q-Networks* para juegos Atari. De

manera similar, Sutton and Barto (2018) destacan que una condición esencial para la estabilidad del aprendizaje es que el estado se exprese en una forma compacta y de dimensión fija. En el ámbito financiero, François-Lavet et al. (2018) en su libro muestran cómo transformar secuencias de precios en ventanas vectorizadas para entrenar agentes de trading basados en redes neuronales.

En cada instante de tiempo discreto $t \in \mathbb{N}$ el agente observa $O_t = \text{vec}(O_t)$ desde $\mathcal{O}$ y toma un vector de decisión d_t en base a su comportamiento (políticas), la observación la cual dispone, y las posibilidades de decisión: 3^N (nueve en este caso). Dadas por:

$$\mathcal{A} = \{0, 1, 2\}^2$$

donde cada componente de la acción $a_t^{(i)}$ se interpreta como:

- $0 \mapsto -1$: vender el activo i (posición corta),
- $1 \mapsto 0$: mantener posición neutra en el activo i,
- $2 \mapsto +1$: comprar el activo i (posición larga).

De esta manera, en cada paso de decisión t el agente observa obs_t, selecciona una acción $a_t \in \mathcal{A}$ y actualiza sus posiciones. El objetivo del agente es aprender una política π que maximice la recompensa terminal del episodio, definida como el índice de Sharpe anualizado de los retornos obtenidos por el portafolio, como se ve a continuación.

4.1.2. Diseño del Sistema de Recompensas

En esta sección se detalla el diseño de la función de recompensa utilizada por el agente del modelo de RL para evaluar sus decisiones, así como la forma en que se estructura el entorno en términos de señales de retorno y aprendizaje. Esta representa una sección muy importante, pues la función de recompensa es el vínculo entre el objetivo financiero y el proceso de aprendizaje del agente. En esta metodología, la señal no se define como premios o castigos locales por aciertos puntuales, sino como una evaluación **episódica**: el agente construye un portafolio a lo largo de un episodio y, sólo al finalizarlo, recibe una recompensa que resume

su desempeño global. En concreto, se utiliza el **índice de Sharpe anualizado** como métrica de desempeño ajustada por riesgo, calculado sobre la secuencia de retornos del portafolio inducida por las decisiones del agente durante el episodio y descontando la tasa libre de riesgo a frecuencia diaria.

El uso del índice de Sharpe obedece a dos razones. Primero, integra en una sola magnitud la *rentabilidad* (media de retornos) y el *riesgo* (volatilidad) Sharpe (1966), alineando el entrenamiento con un criterio ampliamente adoptado en la práctica financiera para comparar estrategias de inversión bajo niveles de incertidumbre distintos. Segundo, la evaluación al final del episodio evita sesgos miopes asociados a recompensas locales: en mercados con ruido y dependencia temporal, una secuencia de decisiones que parezcan "correctas" paso a paso puede no traducirse en un buen desempeño acumulado. Medir el Sharpe *ex post* sobre toda la trayectoria obliga al agente a aprender políticas coherentes en el tiempo, privilegiando estabilidad junto con retorno.

Para medir el retorno del portafolio al final de cada episodio, considere $p_t^{(i)} \in \{-1, 0, +1\}$ la posición del activo i en el paso t, inducida por la acción del agente (venta/posición corta, mantener/neutra, compra/posición larga), y sea $R_{t+1}^{(i)}$ el retorno logarítmico realizado del activo i en el intervalo $[t, t+1]$. El retorno del portafolio en ese intervalo se define como:

$$R_{t+1}^{\text{port}} = \frac{1}{N} \sum_{i=1}^{N} p_t^{(i)} \, r_{t+1}^{(i)}.$$

Este retorno se acumula en una secuencia $\{R_{t_0+1}^{\text{port}}, \ldots, R_{t_0+L}^{\text{port}}\}$ a lo largo del episodio, donde t_0 es el inicio del episodio y L su longitud efectiva. En esta versión, cada activo tiene la misma ponderación en el cálculo del retorno del portafolio, de modo que las diferencias en desempeño provienen únicamente de la dirección de las posiciones (corta, neutra o larga).

Al finalizar el episodio, se calculan los **excesos de retorno diarios** restando la tasa libre de riesgo anualizada[1] a nivel diario ($R_f/252$) a cada retorno del portafolio:

$$\tilde{R}_\ell = R_{t_0+\ell}^{\text{port}} - \frac{R_f}{252}, \qquad \ell = 1, \ldots, L.$$

[1]Por ejemplo, con la TIR de los bonos del tesoro de estados unidos a 10 años

La recompensa terminal es el **índice de Sharpe anualizado** de esta secuencia:

$$G \;=\; \begin{cases} \sqrt{252}\,\dfrac{\overline{\tilde{R}}}{\sigma(\tilde{R})}\,, & \text{si } L > 1 \text{ y } \sigma(\tilde{R}) > 0, \\ 0, & \text{en caso contrario.} \end{cases}$$

donde $\overline{\tilde{R}}$ es la media muestral de los excesos de retorno y $\sigma(\tilde{R})$ su desviación estándar. Durante el episodio, la recompensa intermedia es nula ($g_t = 0$ para todo $t < t_{\text{end}}$), y sólo en el último paso se asigna $g_{t_{\text{end}}} = G$.

En la implementación del entorno, un episodio comienza en un índice inicial t_0 y avanza paso a paso hasta que ocurre alguna de dos condiciones de término. La primera es si se alcanza el límite de datos (no hay más observaciones para avanzar), y la segunda condiciones es si es que se alcanza el horizonte máximo de pasos predefinido (el hiperparámetro llamado `max_steps`). En cualquiera de estos dos casos, el episodio concluye y se genera la recompensa terminal.

Durante el episodio la recompensa intermedia es nula. Esta decisión, coherente con la función `step` del entorno, persigue dos fines. Por un lado, evita entregar señales locales que podrían ser ruidosas o inconsistentes con el objetivo final. Por otro lado, concentra la información en una única métrica global que castiga la volatilidad y premia la estabilidad de los retornos acumulados. Al eliminar el refuerzo paso a paso, el agente no es incentivado a "perseguir" movimientos inmediatos, sino a organizar una secuencia de decisiones cuyo resultado agregado maximice el Sharpe según la recompensa terminal mostrada anteriormente. El costo de esta elección es que la señal de aprendizaje se vuelve más escasa. Sin embargo, ello mantiene alineado el objetivo del entrenamiento con el criterio de evaluación *ex post* utilizado en finanzas.[2]

En consecuencia, el objetivo del agente es encontrar una política π^* que maximice el valor esperado de la recompensa terminal:

$$\pi^* = \arg\max_{\pi} \mathbb{E}[G],$$

donde G es el índice de Sharpe anualizado del portafolio inducido por las decisiones del agente dentro de un episodio.

[2]El criterio ex post es un análisis que se realiza **después** de la ocurrencia de un evento.

4.2. Entrenamiento del Agente con Proximal Policy Optimization (PPO)

Para aproximar esta política óptima π^* definida previamente, se entrena un agente utilizando el algoritmo **Proximal Policy Optimization (PPO)** sobre el entorno personalizado definido en la Sección anterior. PPO fue seleccionado por su equilibrio entre estabilidad y eficiencia. A diferencia de métodos como REINFORCE, que sufren alta varianza, o TRPO, que impone restricciones más costosas computacionalmente, PPO garantiza actualizaciones estables, mediante un término de recorte (*clipping*) que limita el cambio de política, y también mediante la optimización de primer orden con penalización por divergencia KL implícita. Lo que lo convierte en una opción ideal para entornos con decisiones discretas múltiples, como el portafolio de inversión considerado. En nuestro caso, el espacio de acción es de tipo MultiDiscrete, con tres opciones por activo (vender (0), mantener (1) o comprar (2)). En la implementación de la biblioteca utilizada en python `Stable-Baselines3`, la política estocástica de PPO modela este espacio como un producto de distribuciones categóricas independientes, similar a la factorización mostrada en el capítulo 2 de nociones básicas:

$$\pi_\theta(a_t \mid \text{obs}_t) = \prod_{i=1}^{N} \pi_\theta^{(i)}(a_t^{(i)} \mid \text{obs}_t),$$

donde cada factor corresponde a la probabilidad de la acción seleccionada en un activo específico. Esto permite representar de forma compacta decisiones conjuntas sobre múltiples activos, preservando la independencia condicional entre ellos. Este tipo de problema, donde se deben tomar múltiples decisiones simultáneamente sobre distintos activos, emula el proceso de toma de decisiones de un gestor financiero que debe considerar señales recientes de mercado, minimizar errores tácticos y evitar sobreoperar en contextos de baja volatilidad. El algoritmo PPO se adapta naturalmente a esta lógica, pues permite capturar políticas estocásticas ajustadas progresivamente mediante gradiente, priorizando estabilidad y consistencia en entornos de alta dimensión y bajo señal-ruido.

El agente recibe como entrada el estado observado la ventana tamaño $w \times N$ aplanada a vector $\mathbb{R}^{w \cdot N}$, donde w es el tamaño de la ventana temporal y N el número de activos. En esta implementación no se aplica normalización a un rango acotado,

aquí el espacio de observación se define como un rango no acotado[3]. Este vector de observación es procesado por una red neuronal profunda con arquitectura definida como $[256, 256, 128]$, la cual actúa como función de aproximación para la política $\pi_\theta(d_t|O_t)$. Esta política determina una acción por activo que busca maximizar la recompensa acumulada esperada, definida en el final de la sección anterior.

Durante el episodio el entorno entrega recompensa nula y una **recompensa terminal** al finalizar (el SR_X anualizado). Con esta señal escasa, PPO emplea *Generalized Advantage Estimation* (GAE) para la estimación de ventajas $\hat{A}_t$ a partir de retornos mayormente nulos y un pago final, mitigando el problema de asignación de crédito diferida y se distribuye retrospectivamente gracias a GAE Schulman et al. (2016). En particular, el valor de γ regula la propagación de esa señal terminal a lo largo del episodio. El entorno evoluciona entonces al nuevo estado O_{t+1}, y se estima la ventaja relativa de la política actual mediante:

$$\hat{A}_t^{\pi_k} = g_t + \gamma \cdot V(O_{t+1}) - V(O_t)$$

Esta fórmula refleja el valor marginal de haber actuado con la política actual versus la política promedio, ponderado por el factor de descuento γ, que en este contexto representa el grado de orientación al largo plazo en la toma de decisiones de inversión. A partir de esta ventaja, se actualizan los parámetros θ de la política con el objetivo de minimizar una función de pérdida que penaliza desviaciones excesivas entre políticas consecutivas. Esta estrategia busca simular la lógica de rebalanceo prudente de portafolios, basada en evitar ajustes bruscos en la asignación de activos que podrían derivar en costos transaccionales elevados o riesgos innecesarios. La función de pérdida se define como:

$$\mathbf{L}^{\text{PPO}}(\theta) = \mathbb{E}_t\left[\min\left(\frac{\pi_\theta(a_t \mid \text{obs}_t)}{\pi_{\theta_{\text{old}}}(a_t \mid \text{obs}_t)} \cdot \hat{A}_t^{\pi_k},\ \text{clip}\Big(\frac{\pi_\theta(a_t \mid \text{obs}_t)}{\pi_{\theta_{\text{old}}}(a_t \mid \text{obs}_t)}, 1-\varepsilon, 1+\varepsilon\Big)\hat{A}^{\pi_k}\right)\right]$$

En donde:

- $\frac{\pi_\theta(a_t \mid \text{obs}_t)}{\pi_{\theta_{\text{old}}}(a_t \mid \text{obs}_t)}$ es el ratio de probabilidad entre la nueva política y la antigua.
- $\hat{A}_t^{\pi_k}$ es la ventaja estimada mediante GAE para mejorar la estabilidad durante el entrenamiento.

[3]Si bien un activo no podrá perder más de un 100 % de su valor, si podrá aumentar más de un 100 % e inclusive 1000 % entre periodos según frecuencia.

- ε es un parámetro de recorte que controla el cambio de política por cada iteración.

El operador clip se define como:

$$\text{clip}\left(\frac{\pi_\theta(a_t \mid \text{obs}_t)}{\pi_{\theta_{\text{old}}}(a_t \mid \text{obs}_t)}, 1-\varepsilon, 1+\varepsilon\right) = \begin{cases} 1-\varepsilon, & \text{si } \frac{\pi_\theta(a_t \mid \text{obs}_t)}{\pi_{\theta_{\text{old}}}(a_t \mid \text{obs}_t)} < 1-\varepsilon \\ \frac{\pi_\theta(a_t \mid \text{obs}_t)}{\pi_{\theta_{\text{old}}}(a_t \mid \text{obs}_t)}, & \text{si } 1-\varepsilon \leq \frac{\pi_\theta(a_t \mid \text{obs}_t)}{\pi_{\theta_{\text{old}}}(a_t \mid \text{obs}_t)} \leq 1+\varepsilon \\ 1+\varepsilon & \text{si } \frac{\pi_\theta(a_t \mid \text{obs}_t)}{\pi_{\theta_{\text{old}}}(a_t \mid \text{obs}_t)} > 1+\varepsilon \end{cases}$$

Este evita actualizaciones excesivas que podrían volver inestable el proceso de entrenamiento, restringiendo los cambios de política a una zona de confianza. De este modo, PPO evita grandes cambios en la política que podrían desestabilizar el proceso de entrenamiento, asegurando que las nuevas políticas no se desvíen drásticamente de las anteriores mientras se mejora el desempeño (Schulman et al., 2017).

4.2.1. Hiperparámetros Utilizados, Evaluación y Registros del Entrenamiento

La implementación se realizó utilizando la librería `Stable-Baselines3` (Raffin et al., 2021) que ofrece buenas implementaciones de algoritmos de RL. El entrenamiento del agente se llevó a cabo con los siguientes hiperparámetros clave:

La arquitectura de la red neuronal implementada consiste en:

```
net_arch=[256, 256, 128]
```

con activaciones ReLU entre capas. La elección de esta arquitectura responde a la necesidad de capturar no solo patrones **lineales**, sino también relaciones no lineales entre activos, tales como señales cruzadas o cambios de régimen, a fin de buscar modelar o comprender la volatilidad y sus cambios entre fases, como lo expuesto en el capítulo 2. La activación ReLU permite mantener una interpretación financiera sencilla, la cual es respuestas positivas frente a señales relevantes, y saturación

Parámetro	Valor
Tamaño de ventana (w)	10 pasos históricos
Máximo pasos por episodio	265
Tasa de aprendizaje	0.0003
Coeficiente de entropía	0.01
Pasos por actualización (n_steps)	1024
Tamaño de batch	64
Factor de descuento (γ)	0.99
GAE Lambda	0.95
clip_range (ε)	0.2

nula ante movimientos marginales. Asimismo, la normalización de ventajas está activada para estabilizar el entrenamiento y evitar sobreponderar señales falsas de alta varianza.

Durante el entrenamiento, se utilizó un `EvalCallback` con una frecuencia de 5000 pasos para evaluar al agente en un entorno de validación independiente, registrando el desempeño y almacenando automáticamente el mejor modelo encontrado hasta ese punto. Este mecanismo permite monitorear el aprendizaje sin contaminar la fase de entrenamiento, asegurando una comparación objetiva entre distintas iteraciones de la política.

```
eval_callback = EvalCallback(env_val, best_model_save_path="./logs/",
                             log_path="./logs/", eval_freq=5000)
```

Además, se empleó `TensorBoard` para registrar métricas clave de entrenamiento como la recompensa media, la pérdida de valor, la entropía de la política y el ratio de clipping. La visualización de estas curvas resulta esencial para diagnosticar comportamientos no deseados. Pues un descenso prematuro de la entropía puede señalar colapso de exploración; una pérdida de valor persistentemente alta puede indicar dificultad para estimar $V(O_t)$, y un ratio de clipping cercano a cero puede sugerir que las actualizaciones están siendo demasiado agresivas o demasiado conservadoras.

Estas métricas mencionadas son muy utilizadas a la hora de la evaluación durante el entrenamiento de modelos de machine learning, y tienen cierta interpretabilidad

financiera que permiten entender cómo afectan sus valores al resultado obtenido. Para sintetizar, en la Tabla 4.1 se explicita lo mencionado.

Métrica	**Interpretación en entrenamiento financiero**
Recompensa media	Evalúa la capacidad del agente para construir portafolios rentables y estables en el entorno de validación. Un aumento sostenido refleja políticas más efectivas.
Entropía de política	Mide el nivel de exploración. Valores altos indican diversificación de acciones (útil en fases iniciales); valores bajos reflejan mayor confianza en una estrategia concentrada.
Pérdida de valor	Indica qué tan bien la red crítica estima la función de valor $V(O_t)$. Una pérdida elevada puede señalar que el agente tiene dificultades para evaluar correctamente las decisiones de inversión.
Ratio de clipping	Proporción de actualizaciones de política afectadas por el recorte (*clipping*). Valores intermedios indican un entrenamiento estable; valores cercanos a cero pueden implicar actualizaciones poco efectivas, y valores muy altos sugieren cambios de política demasiado agresivos.
Pérdida de política	Refleja la magnitud de los ajustes en la política. Cambios abruptos pueden traducirse en estrategias volátiles; ajustes suaves tienden a generar políticas más consistentes.

Tabla 4.1. Métricas de entrenamiento registradas en `TensorBoard` para PPO

El monitoreo constante de estas señales permite detectar problemas como sobreajuste temprano, estancamiento del aprendizaje o exploración insuficiente, y ajustar el entrenamiento en consecuencia.

4.3. Calibración de Parámetros y Backtesting

En esta sección se desarrolla la metodología utilizada para calibrar los hiperparámetros críticos del modelo y validar su desempeño fuera de muestra. A diferencia de la etapa anterior, centrada en la definición del entorno de entrenamiento y la política aprendida con PPO, aquí se introduce un procedimiento de **backtesting**

que permite evaluar ex-post la calidad de las señales generadas por el agente y seleccionar umbrales de decisión óptimos.

El proceso se estructura en tres componentes principales:

1. Calibración de scores de probabilidad en la fase de validación, que mapea las diferencias $\mathbf{P}(\text{buy}) - \mathbf{P}(\text{sell})$ a retornos esperados y volatilidades condicionales,
2. Estimación de un umbral dinámico τ basado en medidas de riesgo ex-ante mediante un esquema de Expected Shortfall con doble velocidad de memoria,
3. Exploración de una grilla de valores para un parámetro k, que regula la sensibilidad de las decisiones de compra/venta frente al umbral de riesgo, seleccionando aquel que optimiza métricas de desempeño como el índice de Sharpe, el retorno acumulado, la varianza o el Expected Shortfall

La idea de la implementación de una grilla buscando valores de k que optimicen ciertas métricas financieras, es para que el modelo no sólo sea capaz de aprender una política a través de RL, sino también de ajustar sus reglas de decisión mediante la evidencia empírica obtenida en datos de validación.

4.3.1. Calibración del score direccional en validación

Una vez entrenado el agente con el algoritmo de RL descrito en la sección anterior, es necesario evaluar la calidad de las probabilidades que éste asigna a las acciones de compra, mantención o venta. Para ello se recurre a un split de validación independiente de los datos de entrenamiento, lo que permite analizar si la señal producida por la política aprendida mantiene capacidad predictiva fuera de muestra. Este procedimiento cumple un rol análogo al de la *calibración* en modelos de clasificación, donde se busca comprobar hasta qué punto una probabilidad reportada por el modelo se corresponde con una frecuencia empírica observada en los datos. Para desagregar este análisis de los efectos de control del propio agente, se recorre **toda** la ventana del entorno de validación utilizando siempre la acción fija de mantener (no realizar cambio de posiciones).

El agente toma los retornos de la observación y, en base a la estructura y similitud de la observación respecto a experiencias pasadas, en cada estado t el agente produce para cada activo i un vector de probabilidades

$$\big(\mathbf{P}_t^{(i)}(\text{sell}\,|O_t),\ \mathbf{P}_t^{(i)}(\text{hold}\,|O_t),\ \mathbf{P}_t^{(i)}(\text{buy}\,|O_t)\big).$$

Donde el agente busca la opción más conveniente en el sentido de que, si la observación se parece a situaciones en donde comprar obtuvo buenos resultados, entonces la probabilidad de comprar ($P_t^{(i)}(\text{buy}\,|O_t)$)) será más alta.

A partir de estas probabilidades definimos un **score direccional**:

$$s_t^{(i)} = \mathbf{P}_t^{(i)}(\text{buy}\,|O_t) - \mathbf{P}_t^{(i)}(\text{sell}\,|O_t),$$

que resume en un único número la inclinación del modelo hacia posiciones largas (valores positivos) o cortas (valores negativos). Esta construcción es estándar en problemas de predicción de dirección, donde interesa medir el **grado de convicción neta** a favor de un movimiento alcista o bajista.

El objetivo es aproximar una función

$$f_i :\ s \mapsto \mathbf{E}[r^{(i)} \mid s],$$

que describa el retorno esperado del activo i condicionado al score reportado por el agente. Dado que la relación entre s y los retornos futuros puede ser compleja y no lineal, se opta por un enfoque no paramétrico conocido como **binning** o particionamiento por intervalos.

El binning es una técnica utilizada en estadística no paramétrica que consiste en discretizar el rango de una variable continua en intervalos (*bins*) y estimar, dentro de cada bin, una estadística resumen de la variable dependiente. Este enfoque proviene inicialmente de la idea de discretización a partir de los histogramas de Pearson, siendo formalizada en Wand and Jones (1995) como técnica estadística no paramétrica que aproxima funciones de densidad o regresión mediante discretización del dominio continuo. Siendo adoptada al machine learning en Niculescu-Mizil and Caruana (2005) para evaluar la calibración de clasificadores. En nuestro caso, lo que se hace es dividir el rango de scores s en M intervalos definidos por cuantiles:

$$-\infty = b_0 < b_1 < \ldots < b_M = +\infty,$$

de modo que cada bin $[b_{j-1}, b_j)$ contiene aproximadamente la misma cantidad de observaciones. Este criterio evita problemas de bins vacíos en regiones de baja densidad y asegura un uso balanceado de los datos de validación.

El recorrido se implementa iniciando en $t_0 = k - 1$ (ventana llena) y avanzando el entorno con acción `hold`; tras cada avance se registra el retorno realizado $R^{(i)}_{t+1}$. De este modo, cada $s^{(i)}_t$ queda **alineado ex–ante** con el retorno futuro $R^{(i)}_{t+1}$.

Con la serie pareada $\{(s^{(i)}_t,\ R^{(i)}_{t+1})\}_t$ en el *split* de validación, se calibra para cada activo una función no paramétrica f_i mediante *binning* en cuantiles del score. Sea $\{b_j\}_{j=0}^{M}$ el particionado por cuantiles (con M bins); para cada bin $[b_{j-1}, b_j)$ se estima

$$\mu^{(i)}_j \;=\; \frac{1}{|\mathcal{I}_j|}\sum_{t\in\mathcal{I}_j} R^{(i)}_{t+1}, \qquad \sigma^{(i)}_j \;=\; \sqrt{\frac{1}{|\mathcal{I}_j|-1}\sum_{t\in\mathcal{I}_j}\big(R^{(i)}_{t+1}-\mu^{(i)}_j\big)^2}, \quad |\mathcal{I}_j|\neq 1,$$

donde $\mathcal{I}_j = \{\, t \,:\, s^{(i)}_t \in [b_{j-1}, b_j)\,\}$ es el conjunto de tiempos en que el score cayó dentro de dicho intervalo. De esta forma, $\mu^{(i)}_j$ representa el retorno medio observado cuando el modelo entregó un score en ese rango, y $\sigma^{(i)}_j$ mide la volatilidad asociada. La calibración devuelve los bordes de bin, sus centros y los perfiles $\mu^{(i)}$ y $\sigma^{(i)}$, permitiendo evaluar *ex post* el retorno esperado condicional a un score dado. En la práctica, para un score s nuevo se obtiene $\widehat{\mu}^{(i)}(s)$ ubicando s en su bin correspondiente:

$$\widehat{\mu}^{(i)}(s) \;=\; \mu^{(i)}_j \quad \text{si } s \in [b_{j-1}, b_j).$$

Este procedimiento, aplicado **por activo** sobre el conjunto de validación, entrega una relación $s \mapsto \widehat{\mu}$ que se utilizará más adelante para definir reglas de toma de posición basadas en umbrales y para evaluar, vía backtesting, la sensibilidad óptima de dichas reglas.

Así entonces, se tiene el mapeo entre la intensidad de la señal direccional $s^{(i)}_t$ y los retornos futuros del activo. Si el agente asigna una alta probabilidad relativa a comprar, se espera que $s^{(i)}_t$ caiga en un bin con $\mu^{(i)}_j > 0$, lo que sugiere retornos futuros positivos. Por el contrario, scores negativos sostenidos deberían alinearse con bins de retorno medio negativo. De este modo, el binning traduce la salida probabilística de la política en expectativas de retorno más fácilmente interpretables y comparables entre activos.

El uso de **binning** en lugar de un modelo paramétrico (p.ej., regresión lineal) presenta tres ventajas relevantes:

1. No impone supuestos funcionales estrictos sobre la relación $s \mapsto R$, adaptándose a posibles no linealidades, a fin de solucionar algunas brechas explicitadas en el capítulo 2 respecto a la limitancia de la utilización de la suposición de relaciones lineales entre activos y/o sus retornos.
2. Entrega no sólo la media condicional, sino también una estimación de la dispersión condicional $\sigma_j^{(i)}$, útil para medir riesgo.
3. Responde a valores atípicos, ya que los cuantiles aseguran cortes balanceados.

Así entonces, la calibración del score direccional mediante binning sobre el conjunto de validación provee una relación $s \mapsto (\mu, \sigma)$ que servirá más adelante como insumo para definir políticas de decisión basadas en umbrales dinámicos de riesgo y evaluar el desempeño del modelo mediante backtesting.

4.3.2. Estimación de riesgo ex–ante y selección de umbrales

VaR y ES con memoria exponencial (EWMA)

Con el fin de construir un umbral de decisión sensible al riesgo de mercado en cada instante, una pieza clave en el diseño de reglas de trading basadas en umbrales es contar con una medida de riesgo *ex–ante*, capaz de reflejar en cada instante la magnitud potencial de pérdidas extremas en los activos. Para ello se adopta un enfoque inspirado en la metodología **RiskMetrics** desarrollada por *J.P. Morgan* a mediados de los noventa (RiskMetrics Group, 1996), ampliamente difundida tanto en la práctica financiera como en la regulación bancaria. La idea central consiste en asignar **pesos exponencialmente decrecientes** a las observaciones pasadas, de manera que retornos más recientes tengan mayor influencia en la estimación de volatilidad y colas.

Formalmente, sea $R_t^{(i)}$ el retorno del activo i en el tiempo t y $x_t^{(i)} = -R_t^{(i)}$ la pérdida asociada. Para cada fecha t, se define un vector de pesos normalizado

$$w_{t,j} = \frac{(1-\lambda)\,\lambda^{\,t-j}}{\sum_{\ell=0}^{t}(1-\lambda)\,\lambda^{\,t-\ell}} \quad \text{para} \quad j = 0, \ldots, t,$$

donde el parámetro de **decay** $\lambda \in (0, 1)$ es el factor de olvido y gobierna la *persistencia de memoria* valores cercanos a 1 producen curvas más estables pero menos reactivas a cambios recientes; en cambio, valores menores (p.ej. 0.94 como el utilizado) generan estimaciones más sensibles a shocks abruptos en los retornos. Esta formulación es equivalente a un promedio móvil ponderado (*Exponentially Weighted Moving Average, EWMA*), muy popular en la estimación de volatilidad condicional (Jorion, 2007).

Con estos pesos se construye, para cada activo i, la distribución empírica ponderada de pérdidas $\{(x_j^{(i)}, w_{t,j})\}_{j=0}^{t}$. El **Valor en Riesgo** empírico[4] al nivel q se obtiene como el q-cuantil de la distribución ponderada,

$$\mathrm{VaR}_t^{(i)}(q) \;=\; \text{ínf}\left\{ x : \sum_{j:\, x_j^{(i)} \le x} w_{t,j} \;\ge\; q \right\},$$

y el **Expected Shortfall** (ES) como el promedio ponderado en la cola por encima del VaR,

$$\mathrm{ES}_t^{(i)}(q) \;=\; \frac{\sum_{j:\, x_j^{(i)} \ge \mathrm{VaR}_t^{(i)}(q)} w_{t,j}\, x_j^{(i)}}{\sum_{j:\, x_j^{(i)} \ge \mathrm{VaR}_t^{(i)}(q)} w_{t,j}}\,.$$

En implementación, para cada t se ordenan las pérdidas, se acumulan los pesos y se localiza el índice q. El ES se calcula como el promedio ponderado en la cola. Este procedimiento devuelve matrices $\mathrm{VaR}_t^{(i)}$ y $\mathrm{ES}_t^{(i)}$ de dimensión $T \times N$.

Otorga mayor sensibilidad a eventos recientes, lo cual resulta crítico en entornos financieros con *clusters* de volatilidad o cambios de régimen como los mostrados en el capítulo 2 y sobre los activos escogidos en el capítulo 3. Desde un punto de vista teórico, el ES es además una **medida coherente de riesgo** en el sentido de Artzner et al. (1999), superando al VaR en propiedades axiomáticas como la subaditividad. Por ello, en esta tesis se prioriza el uso del ES como base para definir umbrales dinámicos de decisión.

La literatura reciente sostiene la utilización de enfoques EWMA en contextos de gestión de riesgos de corto plazo y calibración dinámica de capital regulatorio (McNeil et al., 2015). En línea con estos estudios, aquí se emplea el esquema EWMA

[4] formulación equivalente a la mostrada en el capítulo 2

no sólo para estimar la volatilidad, sino directamente para aproximar la magnitud de pérdidas extremas a través de VaR y ES, que posteriormente alimentan la construcción de umbrales de trading adaptativos.

En la práctica, la estimación del riesgo condicional a partir de retornos financieros enfrenta un dilema entre *reactividad* y *estabilidad*:

- Si se escoge un λ bajo en la fórmula EWMA, los pesos se concentran en observaciones recientes y la medida de riesgo responde rápidamente a shocks abruptos, pero al costo de una elevada varianza y posible sobreajuste al ruido.
- Si se escoge un λ alto, los pesos se distribuyen más homogéneamente y se obtiene una medida suave y estable en el tiempo, pero que reacciona lentamente ante cambios de régimen o crisis repentinas.

Para balancear ambos extremos, se implementa un **esquema de doble velocidad** (*two-speed ES*). Donde lo que se hace es que se calculan dos secuencias de Expected Shortfall con parámetros distintos (una rápida con $\lambda_{\text{fast}} = 0.85$ y otra lenta con $\lambda_{\text{slow}} = 0.97$) y en cada instante se toma el valor máximo:

$$\sigma_t^{(i)} \;=\; \text{máx}\left\{\text{ES}_{\text{fast},t}^{(i)},\; \text{ES}_{\text{slow},t}^{(i)}\right\}.$$

Buscando que la señal de riesgo capture con rapidez movimientos extremos (gracias a la curva rápida), sin perder la estabilidad de un estimador de horizonte más largo (curva lenta). El uso combinado de medidas de diferente velocidad ha sido explorado en literatura de gestión de riesgo, en particular en variantes de *RiskMetrics* y modelos híbridos de volatilidad (véase RiskMetrics Group, 1996; Christoffersen, 2014).

Un aspecto crítico al construir umbrales dinámicos es evitar la anticipación de información futura. Si en el tiempo t se calculara $\sigma_t^{(i)}$ usando también el retorno R_t, el agente estaría tomando decisiones con información que en la práctica aún no estaba disponible. Para corregirlo, se aplica un *desplazamiento ex–ante*:

$$\tilde{\sigma}_t^{(i)} = \sigma_{t-1}^{(i)}, \qquad \tilde{\sigma}_0^{(i)} = \sigma_0^{(i)},$$

de modo que las decisiones en t sólo utilizan riesgo estimado con datos hasta $t-1$. Este ajuste, garantiza que el backtesting sea realista y no se contaminen los

resultados con información futura.

Desde el punto de vista económico, la combinación de curvas rápida y lenta refleja la lógica de un gestor que desea proteger el portafolio ante shocks súbitos, pero sin sobrerreaccionar ante movimientos aislados. Es equivalente a mantener dos horizontes de monitoreo de riesgo. Uno de corto plazo que alerta tempranamente de posibles pérdidas extremas, y otro de largo plazo que asegura consistencia y evita ajustes innecesarios en escenarios estables. Esta construcción hace que el umbral $\tau_t(k)$ sea más robusto y adecuado para la toma de decisiones secuenciales.

Regla de decisión con umbral dinámico

Una vez calibrada la función $s \mapsto \widehat{\mu}^{(i)}(s)$, el siguiente paso es transformar esa señal en una decisión operativa de inversión. Para ello se introduce un **umbral dinámico** que compara la magnitud de la expectativa de retorno con una medida de riesgo condicional en el mismo instante.

La intuición proviene de la práctica del *trading con umbrales*, ampliamente utilizada en modelos cuantitativos. Donde se explicita que no basta con que una señal sea positiva para justificar abrir una posición larga, sino que la magnitud esperada debe ser suficientemente grande en relación con la incertidumbre asociada. De lo contrario, se corre el riesgo de operar en base a fluctuaciones insignificantes, incurriendo en costos de transacción y aumentando el ruido del portafolio. Esta lógica ha sido formalizada recientemente en la literatura de (De March and Lehalle, 2018), en la cual se muestra cómo estructurar y controlar estrategias que dependen de un umbral sobre una señal, alineando la acción del agente con condiciones de mercado estables y rentables.

Para construir el umbral de decisión, considere $\widehat{\mu}_t^{(i)}$ la media condicional estimada para el activo i en t, obtenida a partir de la calibración del score. El umbral dinámico se define como

$$\tau_t^{(i)}(k) \;=\; k \cdot \tilde{\sigma}_t^{(i)},$$

donde $\tilde{\sigma}_t^{(i)}$ es la medida de riesgo ex–ante derivada del Expected Shortfall (Sección previa), y $k > 0$ es un **parámetro de sensibilidad**. Este parámetro regula cuán exigente es la señal para gatillar una operación. Aquí, valores bajos de k implican

actuar incluso ante señales débiles (mayor frecuencia de operaciones), mientras que valores altos restringen las operaciones a situaciones de convicción fuerte. En términos económicos, k controla el *trade-off* entre aprovechar más oportunidades y reducir costos por sobreoperar.

La política de posiciones persistentes (o regla de decisión) se formula como:

$$p_t^{(i)} = \begin{cases} +1, & \text{si } \widehat{\mu}_t^{(i)} > \tau_t^{(i)}(k), \\ -1, & \text{si } \widehat{\mu}_t^{(i)} < -\tau_t^{(i)}(k) \text{ (si se permite venta corta)}, \\ p_{t-1}^{(i)}, & \text{en caso contrario.} \end{cases}$$

En esta tesis se desactiva la venta corta, por lo que el segundo caso se reemplaza por $p_t^{(i)} = 0$ (lo que significaría vender todo). La persistencia ($p_t^{(i)} \leftarrow p_{t-1}^{(i)}$) introduce histéresis, es decir, si la señal no supera el umbral, se mantienen las posiciones existentes, reduciendo la rotación de activos en las carteras y los costos transaccionales.

Optimización del parámetro k

El parámetro k determina la sensibilidad de la regla de decisión frente a las señales calibradas. Para seleccionarlo de manera objetiva, se plantea como un problema de optimización sobre el conjunto de validación, evaluando métricas de desempeño del portafolio bajo distintos valores de k. Sea $\mathcal{M}(k)$ una métrica de interés (por ejemplo, el índice de Sharpe obtenido al aplicar la política con umbral k), el problema se formula como:

$$k^* \;=\; \arg\max_{k \in \mathcal{K}} \mathcal{M}(k),$$

donde $\mathcal{K}$ es un intervalo compacto, típicamente $\mathcal{K} = [0, 1] \subset \mathbb{R}$. En este marco, pueden definirse varias funciones objetivo:

- **Maximización del Sharpe**:

  $$k^*_{\text{Sharpe}} \;=\; \arg\max_{k \in \mathcal{K}} \text{Sharpe}(k),$$

  donde $\text{Sharpe}(k)$ corresponde al índice de Sharpe anualizado del PnL generado con el umbral k. Este criterio privilegia un balance entre retorno y volatilidad.

- **Minimización del riesgo de cola (Expected Shortfall)**:

$$k^*_{\text{ES}} \;=\; \arg\min_{k\in\mathcal{K}} \text{CVaR}_\alpha(k),$$

donde $\text{CVaR}_\alpha(k)$ es el Expected Shortfall histórico al nivel de confianza α. Este criterio selecciona políticas más conservadoras, enfocadas en limitar pérdidas extremas.

- **Maximización del retorno acumulado**:

$$k^*_{\text{Ret}} \;=\; \arg\max_{k\in\mathcal{K}} \text{RetAcum}(k),$$

centrado en la rentabilidad absoluta sin ajustar por riesgo.

- **Minimización de la varianza**:

$$k^*_{\text{Var}} \;=\; \arg\min_{k\in\mathcal{K}} \text{Var}(k),$$

que privilegia políticas de baja volatilidad, incluso a costa de menores retornos.

Dado que $\sigma_t^{(i)}$ proviene de una estimación de riesgo condicional (Expected Shortfall), multiplicarla por un factor k mayor a 1 no aporta mayor interpretación económica, pues equivaldría a exigir que la señal exceda más de una desviación de riesgo "completa" antes de actuar, lo cual empíricamente anula gran parte de las oportunidades de trading. De forma simétrica, valores negativos de k no tienen sentido pues invertirían la lógica de la regla de decisión, si k fuese negativo, dado que la desviación estándar es no-negativa, entonces se forzaría al modelo a comprar inclusive cuando el retorno esperado sea negativo, lo cual no tiene sentido.

En todos los casos, el dominio $\mathcal{K}$ es un intervalo compacto y las métricas $\mathcal{M}(k)$ se obtienen mediante simulación de backtesting para cada valor de k. Las métricas de desempeño $\mathcal{M}(k)$ (Sharpe, retorno acumulado, varianza, CVaR) se obtienen mediante simulación de backtesting para cada valor de k. Como se trata de funcionales de trayectorias de retornos, no admiten una expresión analítica simple y pueden presentar múltiples extremos locales. Para garantizar que la solución escogida corresponda al óptimo global y no a un punto subóptimo, se implementa una **búsqueda exhaustiva en grilla**, evaluando 1000 valores equiespaciados en $[0, 1]$. De este modo, se identifica el máximo (o mínimo) global observable en el dominio, sin depender de algoritmos de optimización locales que podrían estancarse.

El umbral dinámico $\tau_t^{(i)}(k)$ puede verse como una *barrera de inacción*, en otras palabras, si la expectativa de retorno neta no compensa el nivel de riesgo estimado, se opta por mantener la posición. Lo que es una idea similar al de los *costos de oportunidad y de transacción*, pues operar ante señales débiles genera poca ganancia incremental pero sí costos ciertos (comisiones, spreads, slippage). En cambio, exigir que la señal supere una fracción del riesgo esperado asegura que las operaciones sean justificadas por un ratio señal-ruido atractivo.

Construcción de PnL con costos de transacción

Dadas las posiciones $p_t \in \{-1, 0, 1\}^N$ y los retornos $R_t \in \mathbb{R}^N$, el rendimiento bruto por paso es

$$g_t = \sum_{i=1}^{N} p_t^{(i)} \, R_t^{(i)}.$$

Sin embargo, este valor depende del número de activos en los que el portafolio mantiene una posición activa en ese instante. Para evitar que episodios con distinta cantidad de posiciones abiertas resulten no comparables, se normaliza dividiendo por $K_t = \sum_i |p_t^{(i)}|$, el número de activos en los que efectivamente se está largo o corto. De este modo se obtiene un retorno medio por activo en posición:

$$\tilde{g}_t \;=\; \begin{cases} \dfrac{1}{K_t} \sum_{i:\,|p_t^{(i)}|=1} p_t^{(i)} \, R_t^{(i)}, & \text{si } K_t > 0, \\[2ex] 0, & \text{si } K_t = 0. \end{cases}$$

Esta normalización garantiza que el PnL no aumente artificialmente sólo por diversificar en más activos, sino que refleje un promedio de rendimiento por exposición efectiva.

El costo de transacción se modela como proporcional al *turnover*, entendido como el número de cambios de posición entre dos pasos consecutivos. Intuitivamente, el turnover mide cuántos activos se compran o venden en la transición de $t-1$ a t. Formalmente:

$$\text{turnover}_t \;=\; \sum_{i=1}^{N} \left| p_t^{(i)} - p_{t-1}^{(i)} \right|.$$

Si en un paso se pasa de mantener (0) a comprar ($+1$), o de vender (-1) a mantener, ello incrementa el turnover en una unidad. El costo por paso es entonces:

$$\text{cost}_t \;=\; c \cdot \text{turnover}_t,$$

donde $c > 0$ representa la tasa de costo proporcional por transacción.

Finalmente, el beneficio neto por paso (**PnL**) combina el retorno medio con el costo de transacción:

$$\text{PnL}_t \;=\; \tilde{g}_t \;-\; \text{cost}_t.$$

Métricas de desempeño

Se reportan cuatro métricas sobre la serie $\{\text{PnL}_t\}_{t=1}^{T}$, las cuales son el Sharpe anualizado, el retorno acumulado, la varianza muestral y el expected shortfall al 95 %.

Además, se introduce la **razón de activación promedio**:

$$\text{active\_ratio} \;=\; \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{I}\{|p_t^{(i)}| = 1\},$$

que mide la fracción de pares *(activo, tiempo)* en los que la estrategia mantiene una posición distinta de cero. Esta métrica actúa como indicador de parsimonia, en donde las estrategias con ratios de activación excesivamente altos pueden estar sobreoperando sin necesidad o incurriendo en mayores costos de transacción. En cambio, valores moderados sugieren un balance entre aprovechar señales y mantener la simplicidad operativa.

Selección del parámetro k por backtesting

Sobre el conjunto de validación se evalúa una grilla densa $k \in [0.01, 0.99]$ (paso 0.00098) y se computan las métricas anteriores para cada valor. Se consideran cuatro criterios de selección:

$$k^*_{\text{Sharpe}} \;=\; \arg\max_k \text{Sharpe}(k), \quad k^*_{\text{RetAcum}} \;=\; \arg\max_k \sum_t \text{PnL}_t(k),$$

$$k^*_{\text{Var}} \;=\; \arg\min_{k\in\mathcal{C}} \text{Var}(\text{PnL}(k)), \quad k^*_{\text{ES}} \;=\; \arg\min_{k\in\mathcal{C}} \text{CVaR}_{0.95}(k),$$

donde el conjunto candidato $\mathcal{C}$ filtra soluciones con actividad prácticamente nula. Es decir, que $\mathcal{C} = \{k : \text{active\_ratio}(k) > 0.01\}$. Este filtro evita seleccionar umbrales triviales que induzcan inacción (por ejemplo, posiciones casi siempre nulas) pues así el riesgo sería cero (dada la nula operación y posiciones cero) pero no sería un caso que nos interesara.

Capítulo 5

Benchmarks y Métodos Comparativos

El diseño de un agente de aprendizaje por refuerzo (RL) para la gestión de portafolios exige, además de la calibración de sus parámetros y de la evaluación de su desempeño interno, una comparación rigurosa frente a estrategias alternativas. Este procedimiento permite situar los resultados obtenidos en un contexto más amplio, evitando atribuir mejoras únicamente al modelo propuesto cuando estas podrían ser reproducidas por enfoques más simples o ampliamente aceptados en la literatura financiera.

En consecuencia, este capítulo presenta los **benchmarks y métodos comparativos** que se utilizarán como referencia para evaluar el desempeño del agente entrenado bajo el algoritmo PPO. Los benchmarks cumplen una doble función, primeramente proveen líneas base de distinta complejidad (desde estrategias pasivas como *Buy & Hold* hasta portafolios optimizados bajo el marco de Markowitz), por otro, permiten contrastar el modelo de RL con métodos supervisados y heurísticos que representan aproximaciones alternativas al problema de predicción de retornos y toma de decisiones.

La selección de benchmarks que se utilizan para comparar el modelo de RL, responde a criterios tanto prácticos como metodológicos. En primer lugar, se incluyen enfoques financieros clásicos que constituyen estándares de comparación en estudios de gestión de inversiones. En segundo lugar, se incorporan métodos de predicción estadística y de aprendizaje automático supervisado (Logistic Regression, MLP, XGBoost, LSTM), que permiten evaluar la capacidad del modelo de RL frente a técnicas de predicción de retornos exógenas. Finalmente, se consideran reglas

heurísticas basadas en señales de momentum, ampliamente utilizadas en la práctica por su simplicidad e interpretación directa.

De este modo, los benchmarks que se definen en este capítulo, servirán para valorar ventajas y limitaciones del enfoque propuesto, sirviendo así también como punto de referencia en el análisis de resultados que se desarrollará en el próximo capítulo de esta tesis.

5.1. Benchmarks Financieros Clásicos

Antes de evaluar el desempeño del agente de Aprendizaje por Refuerzo, resulta necesario establecer un conjunto de estrategias de referencia o *benchmarks* que sirvan como línea base. Estos benchmarks representan enfoques tradicionales en finanzas, ampliamente utilizados en la literatura académica y en la práctica de la gestión de portafolios. Su inclusión permite responder a una pregunta fundamental: ¿La política aprendida por el agente de RL aporta un valor agregado real frente a las estrategias más sencillas y conocidas?

En esta sección se consideran tres benchmarks financieros clásicos: **Buy & Hold**, **Portafolio de Markowitz** y **Estrategia de Momentum**. Cada uno de ellos usa una filosofía distinta de inversión:

- **Buy & Hold:** Consistente en adquirir un conjunto de activos y mantenerlos a lo largo de todo el horizonte temporal.
- **Markowitz:** Corresponde al modelo media–varianza clásico, que fundamenta gran parte de la teoría moderna de portafolios.
- **Momentum:** Se trata de una de las anomalías de mercado más estudiadas y se utiliza aquí como un benchmark de tipo heurístico.

Por un lado se tienen estrategias pasivas de baja complejidad (Buy & Hold), por otro, modelos de optimización basados en teoría (Markowitz) y también se tienen reglas empíricas simples pero efectivas (Momentum).

5.1.1. Buy & Hold

La estrategia **Buy & Hold** (comprar y mantener) constituye el punto de partida más elemental para evaluar el desempeño de cualquier estrategia de inversión. Su lógica es directa: se adquiere un portafolio inicial de activos en $t = 0$ y se mantiene sin alteraciones hasta el final del horizonte de inversión.

Desde el punto de vista teórico, Buy & Hold es ampliamente utilizada como benchmark porque refleja el rendimiento intrínseco del mercado, sin intervención activa del gestor. En contextos académicos, su relevancia radica en dos aspectos, la primera y más primordial es que establece una referencia mínima contra la cual debe medirse cualquier estrategia activa. Y, en segundo lugar permite aislar el efecto de las decisiones dinámicas del agente de RL respecto a una política pasiva.

Formalmente, si $R_t \in \mathbb{R}^N$ denota el vector de retornos logarítmicos de N activos en el tiempo t, y $w \in \mathbb{R}^N$ el vector de ponderaciones iniciales del portafolio, entonces el valor acumulado del portafolio bajo Buy & Hold se define como:

$$PV_t^{\text{B\&H}} \;=\; (1-c)\cdot \sum_{i=1}^{N} w_i \cdot \exp\left(\sum_{\tau=1}^{t} R_\tau^{(i)}\right),$$

donde c representa el costo inicial de transacción (comisión por armar el portafolio). En la práctica, cuando no se especifica una preferencia particular, se asume un portafolio igualmente ponderado ($w_i = 1/N$).

5.1.2. Portafolio de Markowitz

El modelo de **Markowitz** (Markowitz (1952)) constituye el fundamento de la teoría moderna de portafolios. Su objetivo es determinar combinaciones de activos que optimicen una métrica de desempeño bajo restricciones de presupuesto y, opcionalmente, límites por activo. Dentro de esta estructura, se definen dos problemas canónicos, el primero es la minimización de la varianza para un nivel dado de retorno, o la maximización del retorno sujeto a restricciones de asignación.

En este trabajo, se utiliza una versión simplificada orientada a la **maximización del retorno esperado**. Dado un conjunto de retornos de entrenamiento $R_t^{(i)}$, se estima la media muestral

$$\mu_i \;=\; \frac{1}{T}\sum_{t=1}^{T} R_t^{(i)},$$

para cada activo i. Luego, la regla de asignación selecciona el activo (o combinación de activos) con mayor media, respetando un límite máximo de concentración (*cap*) en cada activo:

$$w^* \;=\; \arg\max_{w \in \mathcal{W}} \; \mu^\top w,$$

sujeto a $w_i \in [0, \text{cap}]$ y $\sum_i w_i = 1$.

Cuando no se especifica un límite de concentración, la solución degenera en una *cartera concentrada* con todo el peso en el activo de mayor retorno esperado. En cambio, al imponer un cap, se fuerza la inversión de manera secuencial entre los activos de mayor μ_i hasta agotar la masa total.

El benchmark de Markowitz captura una estrategia estática optimizada *ex ante*, basada únicamente en el retorno esperado de los activos. A diferencia de Buy & Hold, no refleja la trayectoria "natural" del mercado, sino una asignación inicial diseñada para maximizar rendimiento bajo un criterio determinista. En consecuencia, este benchmark sirve como límite superior estático, pues cualquier política dinámica, como la del agente de RL, debería compararse no sólo contra la inercia de mantener pesos fijos, sino también contra la asignación más favorable identificada a partir de los datos históricos de entrenamiento.

5.1.3. Estrategia de Momentum

El **momentum** es una de las anomalías más documentadas en las finanzas empíricas ((Jegadeesh and Titman, 1993)). Se basa en la premisa de que los activos con retornos positivos recientes tienden a seguir comportándose bien en el corto plazo, mientras que los de retornos negativos recientes tienden a continuar rezagados. En otras palabras, el momentum aprovecha la persistencia temporal de los retornos en horizontes de corto a mediano plazo.

Formalmente, para cada activo i se define un *retorno acumulado de lookback* en el instante t como:

$$M_t^{(i)} \;=\; \sum_{j=1}^{L} r_{t-j+1}^{(i)},$$

donde L es la ventana de observación (*lookback*). La regla de decisión se establece de

la siguiente manera:

$$p_t^{(i)} = \begin{cases} +1, & \text{si } M_t^{(i)} > \tau, \\ -1, & \text{si } M_t^{(i)} < -\tau \text{ (si se permite venta corta)}, \\ 0, & \text{si } |M_t^{(i)}| \leq \tau, \end{cases}$$

donde τ es un *umbral de neutralidad*, que introduce una banda alrededor de cero para evitar transacciones ante señales marginales.

El benchmark de momentum es una estrategia de trading técnico, basada en información histórica reciente. A diferencia de Buy & Hold y Markowitz, que son estáticos, el momentum es un benchmark dinámico, con rebalanceo continuo de acuerdo al signo y magnitud de los retornos acumulados. Esto lo convierte en un punto de referencia intermedio, es más sofisticado que simplemente mantener o asignar ex ante, pero no utiliza modelos de aprendizaje automático.

Para efectos de esta tesis, el momentum cumple un doble rol, sirve como comparación frente a estrategias técnicas clásicas, y actúa como un "puente" metodológico entre benchmarks financieros simples y los modelos de aprendizaje supervisado o por refuerzo.

5.2. Modelos Supervisados

Además de los benchmarks financieros tradicionales, se consideran **modelos supervisados de aprendizaje automático** como competidores del agente de Reinforcement Learning. La motivación para incluir estos métodos radica en que el problema de anticipar movimientos direccionales de retornos puede plantearse como una tarea de *clasificación binaria*. Donde se busca predecir si el próximo retorno será positivo o negativo a partir de información histórica. Esta perspectiva ha sido explorada en numerosos trabajos de la literatura financiera, donde se busca capturar patrones no triviales en los datos de mercado mediante algoritmos predictivos. En (Ticknor, 2013) se habla respecto al uso de redes neuronales para predicción, siendo muy citada en finanzas con machine learning. También, en (Huck, 2009) se trabajan estrategias predictivas y de selección con machine learning, donde se resalta el uso de métodos supervisados en series financieras.

En esta tesis se consideran cuatro enfoques supervisados de distinta naturaleza y complejidad:

- **Regresión logística (LogReg)**: un modelo lineal interpretable, que sirve como referencia mínima para evaluar la separabilidad de las series de retornos.
- **Perceptrón multicapa (MLP)**: una red neuronal feedforward capaz de capturar relaciones no lineales entre retardos de retornos.
- **XGBoost**: un método de ensamble basado en árboles, eficiente para manejar interacciones no lineales y robusto ante ruido en los datos.
- **LSTM**: una arquitectura recurrente especializada en secuencias temporales, que permite modelar dependencias de largo plazo en los retornos.

Cada uno de estos modelos se entrena de manera **independiente por activo**, utilizando un esquema de ventanas deslizantes de longitud L que transforma los retornos pasados en vectores de características, y definiendo como etiqueta el signo del retorno siguiente. Posteriormente, las probabilidades estimadas de retorno positivo se traducen en decisiones discretas de inversión (comprar, mantener, vender) aplicando un umbral de probabilidad simétrico.

La inclusión de este conjunto diverso de modelos supervisados permite contrastar al agente de RL no solo contra estrategias financieras simples, sino también frente a algoritmos de *machine learning* ampliamente utilizados en la predicción de series financieras, evaluando así si el RL aporta ventajas en escenarios donde técnicas predictivas clásicas ya ofrecen desempeños razonables.

En la siguiente sección se definirán las métricas comparativas empleadas para evaluar estos benchmarks frente al agente de RL, asegurando criterios homogéneos de desempeño.

5.2.1. Regresión Logística

La **regresión logística** es un modelo clásico de clasificación binaria, frecuentemente utilizado en finanzas para predecir la dirección de retornos (Christoffersen and Diebold (2006)). Su objetivo es estimar la probabilidad de que el retorno futuro de un activo sea positivo, dadas ciertas características históricas.

En este benchmark, la variable dependiente se define como:

$$y_t^{(i)} = \begin{cases} 1, & \text{si } r_t^{(i)} > 0, \\ 0, & \text{si } r_t^{(i)} \leq 0, \end{cases}$$

y las variables explicativas corresponden a los L retornos pasados, a los que se agrega opcionalmente la volatilidad local estimada en la misma ventana. Esto genera un vector de características:

$$x_t^{(i)} = \left(r_{t-1}^{(i)}, r_{t-2}^{(i)}, \ldots, r_{t-L}^{(i)},\ \widehat{\sigma}_{t,L}^{(i)}\right),$$

donde $\widehat{\sigma}_{t,L}^{(i)}$ es la desviación estándar de los últimos L retornos, utilizada como proxy de riesgo. Pudiendo también ser una métrica de riesgo de cola, como el expected shortfall en este caso.

De esta forma, un retorno positivo esperado implica `compra` ($+1$), uno negativo esperado implica `venta` (-1, si está habilitada la venta corta), y en caso contrario la posición se mantiene neutra (0).

El benchmark de regresión logística representa un enfoque de **aprendizaje supervisado lineal**, en el que las señales se construyen mediante predicción probabilística de la dirección del retorno. Comparado con Buy & Hold, Markowitz y Momentum, LogReg introduce el uso de un modelo estadístico explícito con entrenamiento ex ante en un conjunto de datos.

La principal ventaja de este benchmark es su simplicidad y transparencia. Pues entrega probabilidades interpretables y coeficientes que permiten cuantificar la influencia de cada predictor. Sin embargo, al ser lineal en los predictores, puede fallar en capturar relaciones no lineales o interacciones complejas entre retornos y volatilidad, limitación que se busca superar con benchmarks posteriores más sofisticados como redes neuronales o modelos no paramétricos.

5.2.2. Perceptrón Multicapa (MLP)

El **Perceptrón Multicapa (MLP)** es un modelo de red neuronal artificial que permite capturar relaciones no lineales entre los retornos pasados y la dirección futura del activo. A diferencia de la regresión logística, que asume un vínculo lineal en el espacio de características, el MLP utiliza capas ocultas con funciones de activación no

lineales, lo que le otorga mayor flexibilidad para aprender patrones complejos. Estudios como Heaton et al. (2017) muestran cómo arquitecturas feedforward pueden capturar patrones complejos en los retornos y mejorar decisiones de inversión, lo que justifica su inclusión como benchmark en este trabajo.

En este benchmark, la tarea de clasificación se define en los mismos términos que en la subsección anterior. Aquí, la variable objetivo es

$$y_t^{(i)} = \mathbf{1}\{R_t^{(i)} > 0\},$$

y las variables explicativas corresponden a los L retornos pasados más la volatilidad local (opcional). Así, el vector de entrada para cada activo es

$$x_t^{(i)} = \big(R_{t-1}^{(i)}, R_{t-2}^{(i)}, \ldots, R_{t-L}^{(i)}, \widehat{\sigma}_{t,L}^{(i)}\big).$$

El procedimiento de entrenamiento se implementa en una función la cual genera un modelo independiente por activo. Cada modelo aplica un escalamiento de los predictores y entrena un clasificador MLP con activación ReLU y función de pérdida logística binaria.

Durante la predicción, se calculan probabilidades de subida ($\mathbf{P}(y = 1)$), y a partir de un umbral (`prob_thresh`), se definen las posiciones discretas para los activos.

El MLP puede capturar relaciones no lineales entre retornos pasados y retornos futuros, lo que lo hace más flexible que LogReg y potencialmente más adaptado a la dinámica no lineal de los mercados financieros.

No obstante, el costo de esta mayor capacidad expresiva es el riesgo de sobreajuste, especialmente en series temporales financieras donde la señal puede estar oculta bajo un alto nivel de ruido. La regularización mediante el parámetro α y el uso de *early stopping* (definido en el máximo número de iteraciones) ayudan a mitigar este problema.

El MLP constituye un comparativo **supervisado no lineal**, diseñado para evaluar hasta qué punto los patrones de retornos pasados contienen información predictiva que pueda ser explotada más allá de las técnicas lineales.

5.2.3. XGBoost

El algoritmo **Extreme Gradient Boosting (XGBoost)** fue introducido por Chen and Guestrin (2016), se trata de un método de ensamble basado en árboles de decisión que ha demostrado un desempeño sobresaliente en tareas de predicción tabular, incluidas aplicaciones financieras. Su fortaleza radica en combinar múltiples modelos débiles (árboles de poca profundidad) entrenados secuencialmente, donde cada árbol corrige los errores de los anteriores mediante el gradiente de la función de pérdida. Esto permite capturar interacciones no lineales y estructuras complejas en los datos.

En este benchmark, al igual que en los modelos previos, la tarea de clasificación consiste en predecir la dirección del retorno:

$$y_t^{(i)} = \mathbf{1}\{R_t^{(i)} > 0\},$$

a partir de un vector de características que incluye los L retornos pasados y, opcionalmente, una medida de volatilidad local:

$$x_t^{(i)} = \big(R_{t-1}^{(i)}, R_{t-2}^{(i)}, \ldots, R_{t-L}^{(i)}, \widehat{\sigma}_{t,L}^{(i)}\big).$$

El procedimiento de entrenamiento se encuentra en una función que entrena un modelo independiente por activo. Se utiliza un `XGBClassifier` con parámetros predefinidos de regularización y control de sobreajuste (`max_depth=3`, `subsample=0.8`, `colsample_bytree=0.8`, entre otros). En caso contrario, el código implementa un mecanismo de respaldo mediante un `RandomForestClassifier`, para la comparabilidad del benchmark.

Durante la predicción, se calcula la probabilidad de subida ($\mathbf{P}(y = 1)$) y se define la posición discreta en función de un umbral de decisión.

XGBoost para este caso da un modelo de referencia y ampliamente utilizado en finanzas, donde los métodos de boosting suelen liderar competencias y pruebas de predicción. Siendo capaz de detectar dependencias no lineales y relaciones complejas entre variables, por lo que es un candidato natural para series financieras con ruido y efectos de interacción.

En comparación con los modelos previos, el XGBoost ofrece un punto intermedio entre la flexibilidad de redes neuronales y la interpretabilidad de los métodos basados

en árboles. Su implementación eficiente y regularización explícita lo hacen especialmente competitivo en contextos donde el tamaño de muestra es limitado, como suele ocurrir en ventanas de entrenamiento de series temporales financieras.

5.2.4. LSTM

Las **Long Short-Term Memory networks (LSTM)** son un tipo de redes neuronales recurrentes (RNN) diseñadas para capturar dependencias de largo plazo en secuencias temporales. A diferencia de las RNN tradicionales, las LSTM introducen compuertas de entrada, salida y olvido que permiten regular el flujo de información, mitigando problemas de *desvanecimiento del gradiente* (Hochreiter and Schmidhuber, 1997). Este diseño las ha convertido en una herramienta ampliamente adoptada en predicción de series temporales, incluidas las series financieras, tal como lo estudian Fischer y Krauss, donde demuestran lo eficiente y la alta capacidad predictiva de estas redes neuronales Fischer and Krauss (2018).

En este benchmark, se entrena un clasificador LSTM independiente por activo. La tarea de clasificación consiste en predecir la dirección del retorno:

$$y_t^{(i)} = \mathbf{1}\{R_t^{(i)} > 0\},$$

a partir de una secuencia de longitud L de retornos pasados, transformada mediante un escalador estandarizado:

$$X_t^{(i)} = \left(R_{t-L}^{(i)}, R_{t-L+1}^{(i)}, \ldots, R_{t-1}^{(i)}\right).$$

El modelo implementado consta de:

- Una capa `LSTM` con tamaño oculto configurable (por defecto 32 unidades).
- Una capa `Dropout` para regularización y evitar sobreajuste.
- Una capa densa final (`Linear`) que produce un logit para la clase positiva.

El entrenamiento se realiza mediante el optimizador Adam y una función de pérdida logística binaria (`BCEWithLogitsLoss`), con esquema de *early stopping* sobre un conjunto de validación interno para prevenir sobreajuste.

El procedimiento de ajustar una LSTM para cada activo, genera:

1. Una secuencia de entrenamiento (X, y) construida a partir de ventanas deslizantes de longitud L.

2. Entrenamiento en GPU (si disponible) durante un máximo de 50 épocas, con *patience* de 5 épocas para detener si no mejora la pérdida de validación.

Luego, se calculan probabilidades de subida usando la función sigmoide aplicada a la salida del modelo. A partir de estas probabilidades se definen decisiones discretas:

$$\text{pos}_t^{(i)} = \begin{cases} 1, & \mathbf{P}(y_t^{(i)} = 1 \mid X_t^{(i)}) > \tau, \\ -1, & \mathbf{P}(y_t^{(i)} = 1 \mid X_t^{(i)}) < 1 - \tau, \quad \text{si se permite short}, \\ 0, & \text{en caso contrario.} \end{cases}$$

El benchmark LSTM es un competidor relevante por dos motivos principales. Primero, introduce una red neuronal capaz de explotar dependencias temporales de mayor longitud que las capturadas por modelos de ventana fija como la regresión logística o el perceptrón multicapa (MLP). Segundo, permite contrastar el desempeño del agente de RL frente a un modelo secuencial de propósito general, usado en la literatura de predicción de series financieras.

De este modo, el benchmark LSTM complementa el set de comparadores al situarse en el extremo más flexible y costoso computacionalmente, ofreciendo un contrapeso respecto a los métodos lineales y de ensamble basados en árboles.

5.3. Métricas Comparativas y Procedimiento de Evaluación

Para evaluar de manera homogénea el desempeño del agente de RL frente a los distintos benchmarks financieros y modelos supervisados descritos previamente, se utilizan métricas cuantitativas ampliamente aceptadas en la literatura de gestión de portafolios y control de riesgo. Estas métricas permiten capturar no solo la rentabilidad, sino también el perfil de riesgo asociado a cada estrategia, entregando así una comparación equilibrada entre competidores.

Métricas de desempeño

Dadas las curvas de valor de portafolio $\{PV_t\}_{t=0}^{T}$, se recuerda la definición del capítulo 1 sobre la secuencia de retornos logarítmicos:

$$R_t = \ln\left(\frac{PV_t}{PV_{t-1}}\right), \qquad t = 1, \ldots, T.$$

A partir de esta serie y bajo la suposición de que los precios de los activos utilizados en esta tesis son positivos, se calculan las siguientes métricas comparativas:

- **Retorno medio**:
$$\overline{R} = \frac{1}{T}\sum_{t=1}^{T} R_t,$$
que refleja la rentabilidad promedio por periodo.

- **Retorno acumulado**:
$$R_{\text{cum}} = \sum_{t=1}^{T} R_t = \ln\left(\frac{PV_T}{PV_0}\right),$$
indicador directo de la creación de valor total sobre el horizonte de inversión.

- **Índice de Sharpe**:
$$S = \frac{\overline{R} - R_f}{\sigma(R)}\sqrt{252},$$
donde R_f es la tasa libre de riesgo diaria y $\sigma(R)$ la desviación estándar de los retornos. Esta métrica captura la eficiencia riesgo-retorno.

- **Varianza de retornos**:
$$\text{Var} = \frac{1}{T-1}\sum_{t=1}^{T}\left(R_t - \overline{R}\right)^2,$$
utilizada como proxy clásica de riesgo.

- **Expected Shortfall (CVaR)**:
$$\text{CVaR}_\alpha = \mathbb{E}\big[\, -R_t \,\big|\, -R_t \geq \text{VaR}_\alpha\big],$$
que mide la pérdida media en la cola α de la distribución, con $\alpha = 0.95$ en este trabajo.

Estas métricas son complementarias, pues mientras el retorno acumulado enfatiza la rentabilidad total, el índice de Sharpe y la varianza capturan la eficiencia ajustada por volatilidad, y el CVaR introduce un enfoque conservador frente a pérdidas extremas.

Procedimiento de evaluación

El proceso de evaluación se implementa de manera sistemática para el agente de RL y para todos los benchmarks, siguiendo los pasos:

1. **Generación de posiciones**: las probabilidades del agente (o las señales de los modelos supervisados) se transforman en posiciones discretas $\{-1, 0, +1\}$ mediante reglas de decisión calibradas en validación.

2. **Cálculo de retornos del portafolio**: dadas las posiciones p_t y los retornos de activos r_t, se obtiene el flujo de ganancias y pérdidas (*PnL*) con ajuste por costos de transacción:
$$\mathrm{PnL}_t = p_t^\top R_t - \mathrm{costos}(p_t, p_{t-1}).$$

3. **Construcción de la curva de valor**: el portafolio evoluciona como
$$PV_t = \prod_{\ell=1}^{t} \exp(\mathrm{PnL}_\ell), \qquad PV_0 = 1.$$
Para B %H se aplica un costo inicial de entrada $(1 - c)$ sobre la combinación lineal de curvas por activo.

4. **Cálculo de métricas**: a cada estrategia j se le asocia el vector
$$M_j = \{\overline{r}_j,\ R_{\mathrm{cum},j},\ S_j,\ \mathrm{Var}_j,\ \mathrm{CVaR}_{0.95,j}\}.$$

5. **Comparación**: los resultados se almacenan en tablas y gráficos comparativos, permitiendo ordenar las estrategias según criterios de riesgo-retorno.

El umbral escalar k se elige en validación maximizando Sharpe o minimizando CVaR (sujeto a una mínima razón de actividad), y se **congela** para la fase de testeo. Todos los competidores (RL y benchmarks) se evalúan con el mismo esquema de costos, anualización y horizontes, a finde obtener comparabilidad.

Capítulo 6

Análisis de Resultados

En este capítulo se presentan los resultados obtenidos a partir de la evaluación empírica del agente de RL y de los benchmarks definidos en el Capítulo anterior. El objetivo es analizar comparativamente el desempeño de las distintas estrategias de inversión bajo criterios de rentabilidad ajustada por riesgo, estabilidad temporal y robustez frente a costos de transacción.

La sección se organiza de la siguiente manera. En primer lugar, se describe el diseño experimental, detallando los *splits* de datos (train, validation, test) y los parámetros finales utilizados en cada modelo. Luego, se reportan las métricas de desempeño en el conjunto de validación, que permiten calibrar el umbral k y seleccionar las variantes más prometedoras del agente. Posteriormente, se presentan los resultados en el conjunto de *test*, que constituyen la evaluación definitiva fuera de muestra. Finalmente, se analizan gráficamente las curvas de valor acumulado de portafolio, junto con tablas comparativas que resumen los indicadores principales: retorno medio, retorno acumulado, índice de Sharpe, varianza y Expected Shortfall (CVaR).

Además, se discuten aspectos cualitativos del comportamiento de los modelos, como la frecuencia de operación (*active ratio*), la estabilidad de las señales de inversión y la sensibilidad a parámetros como los costos de transacción o el horizonte de evaluación. El análisis se complementa con una reflexión sobre las ventajas y limitaciones de cada enfoque, enmarcadas en la literatura de finanzas cuantitativas y aprendizaje automático.

6.1. Diseño Experimental

El análisis empírico se estructura en torno a un **pipeline de evaluación fuera de muestra** compuesto por tres etapas: entrenamiento, validación y prueba. Esta división busca mitigar el riesgo de sobreajuste y obtener conclusiones respecto a la capacidad predictiva de cada estrategia.

Partición de datos

El conjunto de retornos logarítmicos de los activos considerados se divide en tres *splits* temporales:

- **Train**: utilizado para entrenar el agente de RL mediante el algoritmo PPO, así como para estimar parámetros de modelos supervisados (LogReg, MLP, XGBoost y LSTM).
- **Validation**: reservado para calibrar hiperparámetros clave, como el umbral k del agente de RL (definido en la Sección 4.3) o la longitud de ventana L en los modelos supervisados. En este split, el entorno se recorre en modo `hold` para extraer probabilidades y construir la calibración $s \mapsto (\mu, \sigma)$.
- **Test**: se emplea exclusivamente para la evaluación final fuera de muestra, asegurando que ningún dato de esta fase haya sido utilizado en procesos de ajuste o calibración previos.

Parámetros finales

Los hiperparámetros de entrenamiento del agente PPO fueron fijados de acuerdo con la discusión metodológica del Capítulo 4, destacando una tasa de aprendizaje de 3×10^{-4}, un tamaño de lote de 64 y una arquitectura de red neuronal con capas $[256, 256, 128]$. En el caso de los modelos supervisados, cada benchmark fue ajustado con los valores estándar indicados en el Capítulo 5 (e.g., penalización $C = 1.0$ en LogReg, arquitectura $(64, 32)$ en MLP, configuración `max_depth=3`, `subsample=0.8` en XGBoost, y una LSTM de 32 unidades ocultas con regularización `dropout=0.1`).

Procedimiento de evaluación

El procedimiento de evaluación sigue tres pasos:

1. Generación de **curvas de valor de portafolio** (PV_t) a partir de las decisiones de cada estrategia, normalizadas a valor inicial $PV_0 = 1$ y considerando un costo transaccional proporcional de 5 puntos base por cambio de posición.

2. Cálculo de **métricas comparativas** (retorno medio, retorno acumulado, índice de Sharpe, varianza y Expected Shortfall al 95 %) a partir de las series de PnL.

3. Comparación **inter-estrategias**, contrastando al agente de RL contra benchmarks financieros tradicionales (Buy & Hold, Markowitz, Momentum) y modelos supervisados (LogReg, MLP, XGBoost, LSTM).

A diferencia de problemas de clasificación genérica, donde métricas como la curva ROC o el AUC miden la capacidad discriminante del modelo, en aplicaciones financieras el objetivo no es maximizar aciertos en la dirección de retornos, sino optimizar el *desempeño económico del portafolio*. Pues un modelo puede exhibir un AUC alto pero generar pérdidas si las decisiones inducen una mala gestión de riesgo o una alta exposición en contextos de volatilidad extrema.

Por ello, la literatura en finanzas cuantitativas enfatiza métricas de **performance ajustadas por riesgo**, como el índice de Sharpe, la varianza de retornos y medidas de riesgo de cola como el Expected Shortfall. Estas métricas permiten evaluar simultáneamente rentabilidad, estabilidad y vulnerabilidad a pérdidas extremas, dimensiones esenciales en la toma de decisiones de inversión (Lo, 2002; Kolm and Ritter, 2019).

Adicionalmente, el uso de **curvas de valor de portafolio** (PV_t) refleja directamente el crecimiento acumulado del capital bajo cada estrategia, lo cual es una práctica estándar en la evaluación empírica de algoritmos de trading. Trabajos recientes en RL para mercados financieros adoptan este mismo enfoque, (Jiang et al., 2017) evalúan agentes de RL mediante el valor acumulado del portafolio y el índice de Sharpe, mientras que (Fischer and Krauss, 2018) comparan LSTM contra benchmarks usando métricas de rentabilidad y riesgo.

En el caso de los **modelos supervisados secuenciales**, como la red LSTM, es necesario disponer de una ventana completa de al menos L observaciones históricas para

generar la primera predicción. Cada ejemplo de entrenamiento se construye a partir de un vector con L retornos pasados y una etiqueta binaria asociada al retorno siguiente. Esto implica que, para un lookback de 30 días, la primera predicción válida sólo puede realizarse en el día 31, pues antes de ello no existen suficientes observaciones para completar la secuencia de entrada. Tal desfase es inherente a la construcción del dataset y ha sido aplicado, por ejemplo, en Fischer and Krauss (2018), donde los autores utilizan *rolling windows* de 30 retornos para entrenar LSTM en predicción direccional. Aunque en los resultados e imágenes de dicho trabajo no se observa un "gap" explícito, metodológicamente las primeras L observaciones no generan predicciones utilizables.

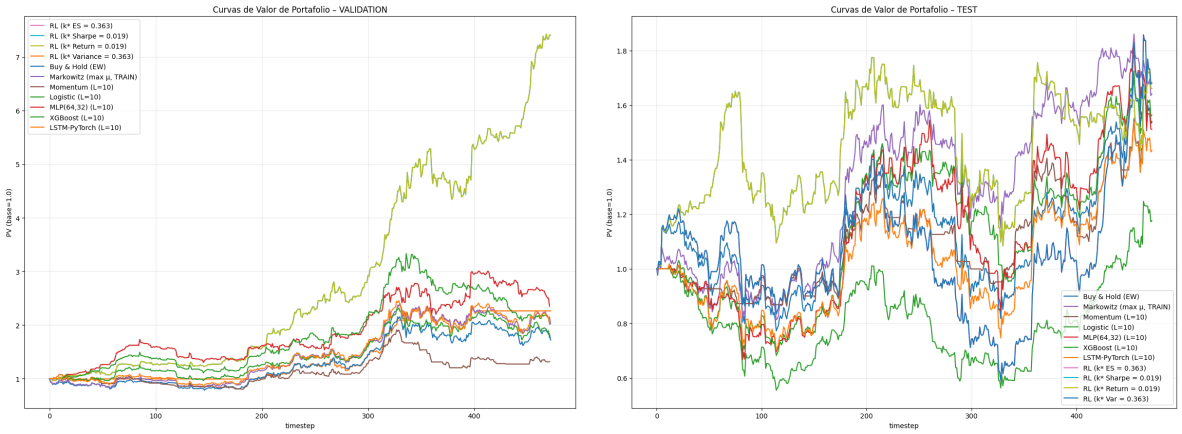
En contraste, el RL en esta tesis no requiere desechar los primeros pasos. El agente recibe como observación un vector de retornos de dimensión fija $w \cdot N$, en caso de no contar aún con suficientes observaciones para completar la ventana, se aplica un **relleno con ceros** (*zero-padding*) hasta alcanzar la dimensión requerida. Cumpliendo así dos propósitos, el primero es asegurar compatibilidad con la definición del espacio de observación del entorno, que debe ser fijo en cada paso para cumplir con la estructura de un proceso de decisión de Markov y la interfaz de los algoritmos de RL, y el segundo es permitir que el agente comience a interactuar con el entorno desde la primera observación disponible, aun cuando la información inicial sea parcial.

La consecuencia práctica es que, en un conjunto de 300 observaciones, una LSTM con $L = 30$ produciría únicamente alrededor de 270 predicciones, mientras que el agente de RL entregará decisiones en los 300 pasos. Visualmente, esto puede hacer que las curvas de desempeño de la LSTM aparezcan "acortadas" respecto a las del RL. Sin embargo, dicha diferencia debe interpretarse como una consecuencia natural de los requerimientos de datos históricos en modelos supervisados secuenciales, y no como una deficiencia metodológica.

Este contraste también explica por qué **no es recomendable rellenar con ceros en el caso supervisado**. Pues allí cada fila del dataset representa un ejemplo etiquetado, y el uso de ceros fabricaría patrones irreales que inducirían shift de distribución y sesgarían el entrenamiento. En RL, en cambio, el padding no se utiliza para construir etiquetas, sino únicamente para entregar al agente un estado inicial válido, si esas observaciones iniciales carecen de información útil, el propio aprendizaje por refuerzo lo penaliza mediante la señal de recompensa, y el agente aprende a actuar de manera conservadora en esas condiciones.

6.2. Resultados sobre Rentabilidad: Criptomonedas

En primera instancia, se utilizó una window size de tamaño 10, que considera las últimas 2 semanas de información por cada paso. Los resultados se resumen en lo siguiente:



(a) Validación (b) Testeo

Figura 6.1. Comparación de curvas de valor de portafolio para la iteración 1.

Estrategia	VALIDACIÓN				TESTEO			
	Ret acum	Sharpe	Var	ES95	Ret acum	Sharpe	Var	ES95
RL (k* ES = 0.363)	0.817	1.506	0.000	0.043	0.520	0.502	0.001	0.075
RL (k* Variance = 0.363)	0.817	1.506	0.000	0.043	0.520	0.502	0.001	0.075
XGBoost (L=10)	0.701	1.170	0.000	0.045	0.162	0.213	0.001	0.060
RL (k* Sharpe = 0.019)	2.002	3.036	0.000	0.045	0.507	0.615	0.001	0.065
RL (k* Return = 0.019)	2.002	3.036	0.000	0.045	0.507	0.615	0.0001	0.065
Momentum (L=10)	0.272	0.468	0.000	0.046	0.430	0.664	0.000	0.050
MLP(64,32) (L=10)	0.855	1.274	0.001	0.053	0.413	0.487	0.001	0.066
Logistic (L=10)	0.567	0.864	0.000	0.053	0.447	0.577	0.001	0.061
LSTM-PyTorch (L=10)	0.703	0.962	0.001	0.054	0.360	0.444	0.001	0.059
Markowitz (max μ, TRAIN)	0.703	0.972	0.001	0.055	0.495	0.686	0.001	0.051
Buy & Hold (EW)	0.542	0.748	0.001	0.056	0.432	0.514	0.001	0.062

Tabla 6.1. Iteración 1: métricas por estrategia.

A partir de las curvas mostradas en la Figura 6.1, en **validación** se observa una dominancia clara de las variantes *RL* calibradas para maximizar retorno/Sharpe: la

trayectoria de *RL (k^* Return = 0.019)* y *RL (k^* Sharpe = 0.019)* crece de forma sostenida, con rupturas de tendencia a favor, y alcanza múltiplos de valor muy superiores al resto. En contraste, en **testeo** las curvas muestran mayor cruce y volatilidad relativa. *RL (k^* Return/Sharpe)* siguen siendo competitivas, pero sin dominancia absoluta. *Markowitz (max μ)* y *Momentum* aparecen más equilibradas, con caídas menos pronunciadas en algunos tramos y recuperaciones rápidas tras los drawdowns.

De las métricas resumidas en la Tabla 6.1, en **validación**, *RL (k^* Return/Sharpe = 0.019)* registran los mejores resultados agregados (retorno acumulado ≈ 2.002, Sharpe ≈ 3.036, ES95 ≈ 0.045), seguidas por *MLP* y *XGBoost*. Las variantes conservadoras *RL (k^* ES/Variance = 0.363)* alcanzan retorno acumulado intermedio (≈ 0.817) pero con un marginal buen control de cola (ES95 ≈ 0.043). En **testeo**, el podio se reordena *Markowitz* obtiene el mayor Sharpe (≈ 0.686) y *RL (k^* Return/Sharpe)* lideran el retorno acumulado (≈ 0.507), aunque con peor ES95 (≈ 0.065) que *Momentum* (≈ 0.050) y *Markowitz* (≈ 0.051). Las versiones *RL (k^* ES/Var)* mantienen un desempeño medio (retorno acumulado ≈ 0.520) pero muestran un deterioro de cola (ES95 ≈ 0.075) respecto de su propio desempeño en validación, e inclusive versus los otros benchmarks. Aunque se mantiene como la mejor en términos de retorno en la fase de testeo.

Comparando fases, las estrategias con umbral agresivo calibrado para retorno/Sharpe (*RL* con k bajo) muestran la mayor *brecha* de desempeño entre validación y testeo (p. ej., ret. acumulado pasa de ≈ 2.002 a ≈ 0.507), lo que sugiere sensibilidad a cambios de régimen y/o sobreajuste del umbral a los patrones de validación. En cambio, los métodos más parsimoniosos o con reglas más estables (*Buy & Hold*, *Markowitz*, *Momentum*, e incluso *Logistic*) presentan caídas más acotadas entre fases y mejor *robustez relativa*. Destaca además que *Momentum* empeora en validación pero mejora en testeo (ret. acumulado $0.272 \rightarrow 0.430$, ES95 $0.046 \rightarrow 0.050$), consistente con un cambio de condiciones de mercado y persistencia de señales.

Desde la estrategia de RL (ver Tabla 6.1), se puede observar un **trade-off explícito**, pues optimizar retorno/Sharpe entrega el mejor crecimiento en validación, pero con mayor degradación en test (más riesgo de cola: ES95 ≈ 0.065). Optimizar ES/Var es más estable en validación (ES95 ≈ 0.043) pero no conserva esa ventaja en test (ES95 ≈ 0.075), lo que indica un desplazamiento del riesgo de cola entre fases. Sin embargo, a pesar de la gran diferencia existente entre ambas fases, la estrategia de trading usando RL se mantiene como la mejor en términos de retorno acumulado.

Bajo una mirada de **perfil de riesgo**, si el trader privilegia retorno, *RL (k* Return/Sharpe)* es competitivo (ret. acum. ≈ 0.507) aceptando mayor cola. Por otro lado, si se privilegia *Sharpe* o protección de cola, *Markowitz* y *Momentum* resultan alternativas más balanceadas (ES95 ≈ 0.051 y ≈ 0.050, respectivamente). + Bajo una mirada de **perfil de riesgo**, si el trader privilegia retorno, *RL (k* Return/Sharpe)* es competitivo (ret. acum. ≈ 0.507) aceptando mayor cola, tal como se aprecia en la Figura 6.1 y la Tabla 6.1. Por otro lado, si se privilegia *Sharpe* o protección de cola, *Markowitz* y *Momentum* resultan alternativas más balanceadas (ES95 ≈ 0.051 y ≈ 0.050, respectivamente).

Segunda iteración sobre criptomonedas

Como segunda iteración, se usa el mismo par de activos pero con la modificación de que ahora la window size corresponde a 20. Quiere decir, un mes completo. En cuyo caso, los resultados son los mostrados en las Figuras 6.2(a) y 6.2(b).

(a) Validación *(b) Testeo*

Figura 6.2. Comparación de curvas de valor de portafolio para la iteración 2.

En **validación**, como se observa en la Figura 6.2(a), las cuatro variantes de *RL* (ES, Sharpe, Var y Return) muestran trayectorias fuertemente crecientes y prácticamente solapadas, superando ampliamente a los benchmarks; el crecimiento es sostenido y con rupturas de tendencia a favor. En **testeo**, mostrado en la Figura 6.2(b), aparece un cambio de régimen con un tramo de corrección intermedio y posterior recuperación; visualmente, varias estrategias (p. ej., *LSTM* y *Markowitz*) alcanzan niveles

de PV elevados hacia el final, mientras que las políticas de *RL* mantienen un perfil competitivo pero con mayor dispersión entre sí.

Estrategia	VALIDACIÓN				TESTEO			
	Ret acum	Sharpe	Var	ES95	Ret acum	Sharpe	Var	ES95
RL (k* ES = 0.034)	1.939	3.588	0.000	0.038	0.610	0.891	0.001	0.060
RL (k* Sharpe = 0.038)	1.978	3.662	0.000	0.039	0.552	0.806	0.001	0.060
RL (k* Variance = 0.037)	1.966	3.645	0.000	0.038	0.530	0.776	0.001	0.060
RL (k* Return = 0.084)	1.980	3.538	0.001	0.043	0.625	0.805	0.001	0.070
MLP(64,32) (L=20)	0.554	1.139	0.000	0.046	-0.069	-0.105	0.001	0.061
Momentum (L=20)	0.478	0.955	0.000	0.050	0.702	1.262	0.001	0.046
XGBoost (L=20)	0.378	0.698	0.000	0.051	0.281	0.465	0.001	0.058
Logistic (L=20)	-0.100	-0.196	0.000	0.053	0.201	0.331	0.001	0.056
Markowitz (max μ, TRAIN)	0.705	1.148	0.001	0.056	0.583	0.959	0.001	0.051
Buy & Hold (EW)	0.593	0.962	0.001	0.057	0.500	0.708	0.001	0.063
LSTM-PyTorch (L=20)	0.591	0.943	0.001	0.058	0.467	0.705	0.001	0.058

Tabla 6.2. Iteración 2: métricas por estrategia.

También, en validación según la Tabla 6.2, la familia *RL* domina, pues *RL–Return* y *RL–Sharpe* registran los mayores retornos acumulados (≈ 1.980 y 1.978) y los mayores índices de Sharpe (≈ 3.54–3.66), con colas acotadas ($ES_{95} \approx 0.038$–0.043). *RL–ES* y *RL–Var* quedan apenas por detrás y con ES_{95} ligeramente menor (≈ 0.038). El resto de modelos queda claramente rezagado. Para **Testeo**, tal como sucede en la iteración 1 y se resume en la Tabla 6.2, el ranking se reordena. *Momentum* lidera en riesgo–retorno con un Sharpe ≈ 1.262 y el mejor $ES_{95} \approx 0.046$ con retorno acumulado ≈ 0.702. Por otro lado, *RL–ES* y *RL–Return* mantienen retornos competitivos (≈ 0.610 y 0.625) pero con colas más pesadas ($ES_{95} \approx 0.060$–0.070) implicando un mayor riesgo de potenciales pérdidas en caso de eventos adversos. *Markowitz* entrega un buen equilibrio (ret. acum. ≈ 0.583, Sharpe ≈ 0.959), *Buy&Hold* queda por detrás (ret. acum. ≈ 0.500), y los modelos supervisados (*LSTM*, *XGBoost*, *MLP*, *Logistic*) muestran desempeño medio a bajo en comparativa.

La brecha validación–test es mayor en las políticas *RL* orientadas a crecimiento (*RL–Return/Sharpe*), que pasan de Sharpe ≈ 3.6 en validación a ≈ 0.80–1.01 en test con ES_{95} superiores. Gran parte de esta brecha existente en las estrategias de RL, se ven argumentadas principalmente por la naturaleza de la metodología. En validación se recorren los posibles τ y se elige el mejor en razón a algún criterio.

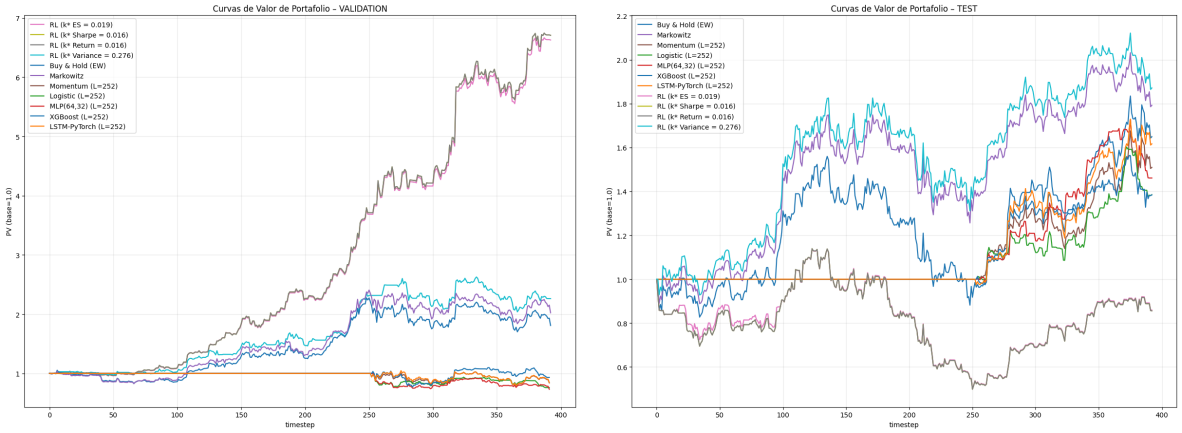
Siguiendo con las estrategias de RL, las variantes *RL–ES/Var* preservan mejor el con-

trol de cola en validación, pero en test no retienen esa ventaja frente a *Momentum*. En este par de activos y ventana (véase Tabla 6.2), **Momentum** resulta la estrategia más *robusta* fuera de muestra (mejor Sharpe y mejor ES_{95}), mientras que **RL** sigue siendo competitivo en retorno acumulado, a costa de mayor riesgo de cola.

Finalizando los comentarios de esta iteración, la evidencia integrada confirma que, con $L = 20$, las políticas de *RL* son muy eficientes en validación y mantienen retornos competitivos en test, pero la estrategia **Momentum** entrega el *mejor equilibrio* riesgo–retorno fuera de muestra (mayor Sharpe y menor ES_{95}). La elección de política depende del objetivo: *crecimiento* (priorizar *RL*) frente a *estabilidad en colas* (*Momentum*/*Markowitz*).

Tercera iteración sobre criptomonedas

Los resultados de la iteración 3 (la final para criptomonedas, con $w = 252$) se presentan en la Figura 6.3, donde se observan las curvas de valor de portafolio en validación y testeo.



(a) Validación *(b) Testeo*

Figura 6.3. Comparación de curvas de valor de portafolio para la iteración 3.

En **validación**, tal como se aprecia en la Figura 6.3(a), las políticas de *RL* orientadas a crecimiento (**RL–Sharpe** y **RL–Return**) y la versión **RL–ES** muestran un avance casi monótono y muy superior al resto, con escalones de ruptura a favor. **RL–Var** también crece, aunque con pendiente menor. Los modelos supervisados (Logistic, MLP, XGBoost, LSTM) y *Momentum* quedan rezagados. En **testeo**, mostrado en la

Figura 6.3(b), el régimen cambia, pues las curvas líderes pasan a ser **RL–Var** y **Markowitz**, que capturan bien los tramos de tendencia y cierran en los niveles de PV más altos. Los métodos supervisados (**LSTM** y **Momentum**) exhiben trayectorias más suaves y estables. En contraste, **RL–Sharpe/Return/ES** pierden tracción y terminan con rendimiento negativo.

Estrategia	VALIDACIÓN				TESTEO			
	Ret acum	Sharpe	Var	ES95	Ret acum	Sharpe	Var	ES95
RL (k* ES = 0.019)	1.891	3.541	0.000	0.034	-0.152	-0.224	0.001	0.068
RL (k* Sharpe = 0.016)	1.903	3.564	0.000	0.034	-0.156	-0.231	0.001	0.068
RL (k* Return = 0.016)	1.903	3.564	0.000	0.034	-0.156	-0.231	0.001	0.068
Logistic (L=252)	-0.274	-0.858	0.000	0.039	0.326	1.103	0.000	0.025
MLP(64,32) (L=252)	-0.310	-0.971	0.000	0.040	0.379	1.192	0.000	0.025
RL (k* Variance = 0.276)	0.817	1.580	0.000	0.044	0.627	1.034	0.001	0.050
XGBoost (L=252)	-0.069	-0.166	0.000	0.045	0.325	0.964	0.000	0.029
LSTM-PyTorch (L=252)	-0.166	-0.379	0.000	0.048	0.480	1.299	0.000	0.030
Momentum (L=252)	-0.174	-0.397	0.000	0.048	0.412	1.272	0.000	0.030
Markowitz	0.705	1.148	0.001	0.056	0.583	0.959	0.001	0.051
Buy & Hold (EW)	0.593	0.962	0.001	0.057	0.500	0.708	0.001	0.063

Tabla 6.3. Iteración 3: métricas por estrategia.

Según los resultados presentados en la Tabla 6.3, para validación **RL–Return** y **RL–Sharpe** lideran (ret. acum. ≈ 1.903, Sharpe ≈ 3.564, $\text{ES}_{95} \approx 0.034$), junto con **RL–ES** (ret. acum. 1.891, Sharpe 3.541). **RL–Var** también es positivo (ret. acum. 0.817, Sharpe 1.580). Los demás modelos muestran retornos negativos en esta fase. Para **testeo**, tal como en las 2 iteraciones anteriores y como se evidencia en la Tabla 6.3, el *ranking* se reorganiza, pues **RL–Var** alcanza el mayor retorno acumulado entre las políticas de *RL* (≈ 0.627, Sharpe ≈ 1.034, $\text{ES}_{95} \approx 0.050$) y **Markowitz** también es fuerte (≈ 0.583, Sharpe 0.959). Los mejores **Sharpes** fuera de muestra los obtienen **LSTM** y **Momentum** (≈ 1.299 y 1.272) con retornos sólidos (≈ 0.480 y 0.412) y colas más contenidas ($\text{ES}_{95} \approx 0.030$). **Buy&Hold** queda por detrás (ret. acum. ≈ 0.500, $\text{ES}_{95} \approx 0.063$). **RL–ES/Sharpe/Return** se tornan negativos (ret. acum. ≈ -0.152 a -0.156) quedando como las peores en desempeño.

Se observa nuevamente una brecha pronunciada entre validación y test, fenómeno que puede interpretarse como sobreajuste al régimen de calibración. Este patrón ha sido documentado en trabajos previos de RL en mercados financieros (como el de Jiang et al. (2017)). En particular, respecto a esta brecha, para *RL–ES/Sharpe/Return*,

pues pasan de Sharpe ≈ 3.55 en validación a valores negativos en test, con ES_{95} más alto (≈ 0.068). Esto indica y evidencia un sobreajuste al régimen de validación y/o subreactividad por la ventana anual ($L = 252$). Por el contrario, **RL–Var** generaliza mejor. Es decir, aunque no domina en validación, en test combina buen retorno con Sharpe positivo. Respecto a los otros benchmarks, la **LSTM** y **Momentum** muestran la mejor eficiencia riesgo–retorno fuera de muestra (mayor Sharpe y colas más livianas), a costa de un retorno absoluto inferior al de **RL–Var**/**Markowitz**. Siendo así distintas estrategias según el perfil de riesgo que tenga el trader. En pocas palabras, se pueden jerarquizar y clasificar los resultados según una **Selección por objetivo:**

- *Crecimiento con aceptación de cola:* **RL–Var** / **Markowitz**.
- *Mayor estabilidad de cola y Sharpe:* **LSTM** / **Momentum** ($\text{ES}_{95} \approx 0.030$).

El aumentar en gran cantidad la ventana (desde 20 en la iteración anterior a 252 en la actual) reduce ruido pero puede demorar ajustes de riesgo, aquellas políticas enfocadas en maximizar retorno/Sharpe (*RL–Return/Sharpe*) quedan expuestas a cambios bruscos de régimen. Una solución a ello, sería reducir L o ajustar k/umbrales para evitar periodos prolongados con exposición inadecuada, incorporando reglas de reentrada y límites de *no-trade* o un overlay pasivo (*EW*) para mejorar la resiliencia.

Los resultados obtenidos tanto para la iteración 1 y 2 son relativamente similares, las estrategias de RL funcionan bien para cuando se dispone de poca observación y/o se requiere de un menor costo computacional. Sin embargo, para w o $L = 252$, las políticas **RL–ES/Sharpe/Return** son sobresalientes en validación pero no generalizan; **RL–Var** y **Markowitz** lideran el retorno en test, mientras que **LSTM** y **Momentum** optimizan el Sharpe y controlan mejor el riesgo de cola. La elección de política debe reflejar la preferencia entre *maximizar retorno* y *minimizar colas* en un entorno altamente cambiante como cripto.

Es importante mencionar también que en esta tesis se abarca el período comprendido entre el **30 de junio de 2018 y el 31 de agosto de 2025**, lo cual revela que el desempeño de las estrategias de inversión estuvo fuertemente influenciado por los cambios de régimen del mercado de criptomonedas.

Periodo de Validación: Esta fase, que va desde **finales de 2022 hasta mediados de 2024**, coincidió con una fase de consolidación y crecimiento gradual en el mercado.

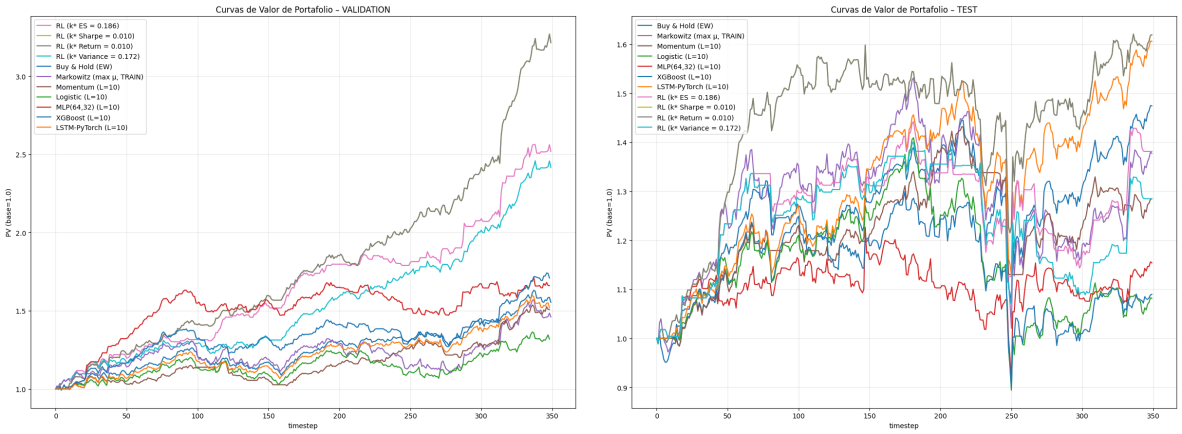
Tras la corrección de 2022, el mercado se estabilizó y entró en un periodo de recuperación, lo que explica por qué las políticas de RL, optimizadas para maximizar el retorno y el **Ratio de Sharpe**, mostraron un rendimiento sobresaliente. Estas estrategias aprendieron a explotar las tendencias de crecimiento de forma muy efectiva.

Periodo de Testeo (Test): A partir de **mediados de 2024 hasta la fecha final del análisis**, el mercado experimentó una volatilidad significativa y un invierno cripto, con fuertes correcciones y periodos de baja. Este cambio brusco de régimen expuso la principal debilidad de las políticas de RL agresivas, la incapacidad para generalizar fuera de las condiciones de mercado en las que fueron entrenadas, dado los cambios de régimen.

La **brecha de rendimiento** observada entre las fases de validación y testeo en las estrategias de RL más agresivas se explica por este cambio de régimen y posible sobreajuste o sobrereacción. Mientras que en validación el entorno favorecía un enfoque de crecimiento, en testeo la volatilidad y las caídas repentinas penalizaron severamente a los portafolios con alta exposición. En contraste, las estrategias más tradicionales como **Momentum** y **Markowitz** demostraron ser también una opción robusta, navegando con mayor eficacia la incertidumbre del mercado. Esto subraya la importancia de la resiliencia y el control de riesgo en un entorno de inversión altamente volátil.

6.3. Resultados sobre Rentabilidad: Acciones

Siguiendo el mismo flujo que para las iteraciones sobre las criptomonedas, para $w = 10$ se obtuvo el conjunto de resultados que se muestran en la Figura 6.4, donde se comparan las trayectorias de valor de portafolio en validación y testeo.



(a) Validación *(b) Testeo*

Figura 6.4. Comparación de curvas de valor de portafolio para la iteración 4.

Estrategia	VALIDACIÓN				TESTEO			
	Ret acum	Sharpe	Var	ES95	Ret acum	Sharpe	Var	ES95
RL (k* ES = 0.186)	0.924	4.357	0.000	0.013	0.323	0.946	0.000	0.035
RL (k* Variance = 0.172)	0.882	4.504	0.000	0.015	0.251	0.733	0.000	0.036
XGBoost (L=10)	0.536	2.479	0.000	0.020	0.086	0.254	0.000	0.039
RL (k* Sharpe = 0.010)	1.167	5.238	0.000	0.017	0.482	1.443	0.000	0.035
RL (k* Return = 0.010)	1.167	5.238	0.000	0.017	0.482	1.443	0.000	0.035
Momentum (L=10)	0.411	1.945	0.000	0.020	0.251	0.993	0.000	0.029
MLP(64,32) (L=10)	0.507	2.326	0.000	0.020	0.143	0.407	0.000	0.035
Logistic (L=10)	0.277	1.185	0.000	0.023	0.079	0.233	0.000	0.038
LSTM-PyTorch (L=10)	0.419	2.074	0.000	0.018	0.474	1.492	0.000	0.032
Markowitz (max μ, TRAIN)	0.379	1.285	0.000	0.028	0.320	0.760	0.000	0.043
Buy & Hold (EW)	0.441	2.100	0.000	0.020	0.388	1.198	0.000	0.034

Tabla 6.4. Iteración 4: métricas por estrategia.

A partir de las curvas mostradas en la Figura 6.4, se observa que en **validación** las políticas de RL calibradas para maximizar el retorno o el Sharpe ($k^* = 0.010$) dominan, pues sus trayectorias son crecientes y alcanzan múltiplos de valor muy superiores al resto, seguidas por *RL–ES* y *RL–Var*, que muestran perfiles más suaves y

persistentes. En **testeo**, la jerarquía cambia, pues las curvas presentan más cruces y un evento de corrección común (t≈250). Tras ese shock, *RL–Sharpe/Return* y *LSTM* se recuperan con fuerza y cierran entre los mejores valores; *Momentum* mantiene una trayectoria más sobria con drawdowns contenidos; *Markowitz* acusa una caída más profunda en el shock; y los modelos supervisados simples (XGBoost, MLP, Logistic) quedan más rezagados.

A partir de la Tabla 6.4, en **validación** se observan que **RL–Sharpe/Return** logran los mejores agregados (retorno acumulado ≈ 1.167, Sharpe ≈ 5.238) con $\text{ES}_{95} \approx 0.017$. Además, **RL–ES** y **RL–Var** presentan retornos acumulados altos (≈ 0.924 y 0.882) y *muy buen control de cola* ($\text{ES}_{95} \approx 0.013$ y 0.015). Por otro lado, el resto de benchmarks queda por detrás (p. ej., *Buy&Hold*: ret. acum. ≈ 0.441, Sharpe ≈ 2.100).

En **testeo**, según la Tabla 6.4, la **LSTM** obtiene el mayor Sharpe (≈ 1.492) y un retorno acumulado competitivo (≈ 0.474). Por otro lado, **RL–Sharpe/Return** lideran en retorno acumulado (≈ 0.482) con Sharpe ≈ 1.443, aunque con ES_{95} algo más alto (≈ 0.035). La estrategia de **Momentum** muestra el *mejor* ES_{95} (≈ 0.029) y un Sharpe cercano a 1 (≈ 0.993), evidenciando *resiliencia* ante el shock. Los benchmark de **RL–ES/Var** bajan en el ranking fuera de muestra (ret. acum. ≈ 0.323 y 0.251; $\text{ES}_{95} \approx 0.035$–$0.036$). Finalmente, **Markowitz** presenta el peor ES_{95} en test (≈ 0.043), consistente con su caída visual más profunda en el evento común.

Las políticas **RL–Sharpe/Return** exhiben la mayor *brecha* entre validación y test (de ret. acum. ≈ 1.167 a ≈ 0.482), sugiriendo sensibilidad al cambio de régimen en el tramo fuera de muestra. **RL–ES/Var** mantienen buen desempeño en validación (colas contenidas), pero en test su ventaja de ES se diluye frente a *Momentum/Buy&Hold*, pero sigue siendo una buena opción, puesa el riesgo de cola es bajo. **LSTM** mejora su posición relativa fuera de muestra, lo que apunta a una mejor capacidad de generalización sobre este par de activos en este período. El conjunto indica que el episodio de estrés en t≈250 reorganiza el ranking y pone en valor estrategias con *riesgo de cola más bajo*.

Como observación final, se puede decir que, bajo los resultados mostrados en la Figura 6.4 y la Tabla 6.4, se confirma el *trade-off* esperado de las iteraciones anteriores: las políticas agresivas de RL maximizan el crecimiento cuando el régimen es estable (validación) y siguen siendo competitivas en retorno fuera de muestra, mientras que estrategias con mejor control de cola (Momentum) y modelos secuenciales (LSTM) resultan más robustos frente al shock observado. La elección de política debiese, por

tanto, ajustarse al perfil del inversor: crecimiento con mayor cola (*RL–Sharpe/Return*) versus protección de cola/estabilidad (*Momentum/LSTM*).

Segunda iteración sobre las acciones

Los resultados para la quinta iteración total (segunda sobre acciones) se resumen en la Figura 6.5, donde se comparan las trayectorias de valor de portafolio en validación y testeo.

(a) Validación *(b) Testeo*

Figura 6.5. Comparación de curvas de valor de portafolio para la iteración 5.

Estrategia	VALIDACIÓN				TESTEO			
	Ret acum	Sharpe	Var	ES95	Ret acum	Sharpe	Var	ES95
RL (k* ES = 0.946)	0.052	1.278	0.000	0.000	0.000	0.000	0.000	0.000
RL (k* Variance = 0.946)	0.052	1.278	0.000	0.000	0.000	0.000	0.000	0.000
RL (k* Sharpe = 0.016)	0.927	5.926	0.000	0.014	0.308	1.218	0.000	0.034
RL (k* Return = 0.010)	0.936	5.926	0.000	0.014	0.299	1.007	0.000	0.037
MLP(64,32) (L=20)	0.169	1.177	0.000	0.018	0.025	0.097	0.000	0.033
Buy & Hold (EW)	0.273	1.784	0.000	0.020	0.266	1.015	0.000	0.036
Momentum (L=20)	0.122	0.731	0.000	0.024	0.181	0.976	0.000	0.027
XGBoost (L=20)	0.297	1.759	0.000	0.022	0.114	0.410	0.000	0.037
Logistic (L=20)	0.045	0.268	0.000	0.024	-0.109	-0.444	0.000	0.034
LSTM-PyTorch (L=20)	0.187	0.822	0.000	0.028	0.060	0.185	0.000	0.045
Markowitz (max μ, TRAIN)	0.140	0.593	0.000	0.030	0.060	0.178	0.000	0.047

Tabla 6.5. Iteración 5: métricas por estrategia.

A partir de la Figura 6.5, se observa que en **validación** las políticas de *RL* orientadas a crecimiento (**RL–Sharpe** y **RL–Return**) muestran una senda marcadamente alcista y

muy por encima de los benchmarks. En cambio, **RL–ES/Var** mantienen un perfil casi plano (exposición baja), mientras que el resto de modelos muestra pendientes menores y trayectorias más irregulares. En **testeo**, varias estrategias capturan el rebote posterior al shock intermedio: **Buy&Hold** y **RL–Return/Sharpe** cierran en los niveles de PV más altos, mientras que **Momentum** evoluciona con oscilación contenida y **RL–ES/Var** permanecen prácticamente inactivas (línea casi horizontal) posiblemente por la baja señal en las probabilidades del RL en contraste con el alto umbral existente. Quiere decir, que como el umbral es tan alto, el criterio de transacción o de cambio de posiciones obedece a que la señal de subida o bajada debe ser fuerte, a fin de disminuir el posible riesgo en tomar una decisión errónea. Por lo cual, el modelo en conjunto con el umbral de decisión opta por tomar una posición más neutral al respecto.

Según la Tabla 6.5, en **validación RL–Sharpe** y **RL–Return** lideran ampliamente (ret. acum. ≈ 0.927–0.936; Sharpe ≈ 5.93; $\mathrm{ES}_{95} \approx 0.014$). **RL–ES/Var** avanzan poco (ret. acum. ≈ 0.052, Sharpe ≈ 1.28). Entre los benchmarks, destacan **Buy&Hold** y **XGBoost** (Sharpe ≈ 1.76–1.78). En **testeo**, el *ranking* se reconfigura: **RL–Sharpe** alcanza el mejor Sharpe fuera de muestra (≈ 1.218) con buen retorno (ret. acum. ≈ 0.308, $\mathrm{ES}_{95} \approx 0.034$). **RL–Return** y **Buy&Hold** siguen de cerca (0.299 y 0.266; Sharpe ≈ 1.01), con colas similares (≈ 0.036–0.037). **Momentum** ofrece el *mejor control de cola* ($\mathrm{ES}_{95} \approx 0.027$) y Sharpe ≈ 0.976, aunque con menor retorno (0.181). **RL–ES/Var** permanecen sin operar (ret. acum. $= 0$). Entre los modelos supervisados, **XGBoost** y **MLP** son positivos pero discretos; **Logistic** y **LSTM** resultan débiles (Sharpe bajo y ES_{95} más alto en *LSTM*).

La **brecha validación–test** es mayor en las políticas orientadas a crecimiento (*RL–Return/Sharpe*), que pasan de Sharpe ≈ 5.9 en validación a ≈ 1.0–1.2 en test y con colas algo más pesadas. Las variantes *RL–ES/Var* muestran *exceso de conservadurismo* y quedan prácticamente planas fuera de muestra. **Momentum** y **Buy&Hold** presentan perfiles más estables en el cambio de régimen.

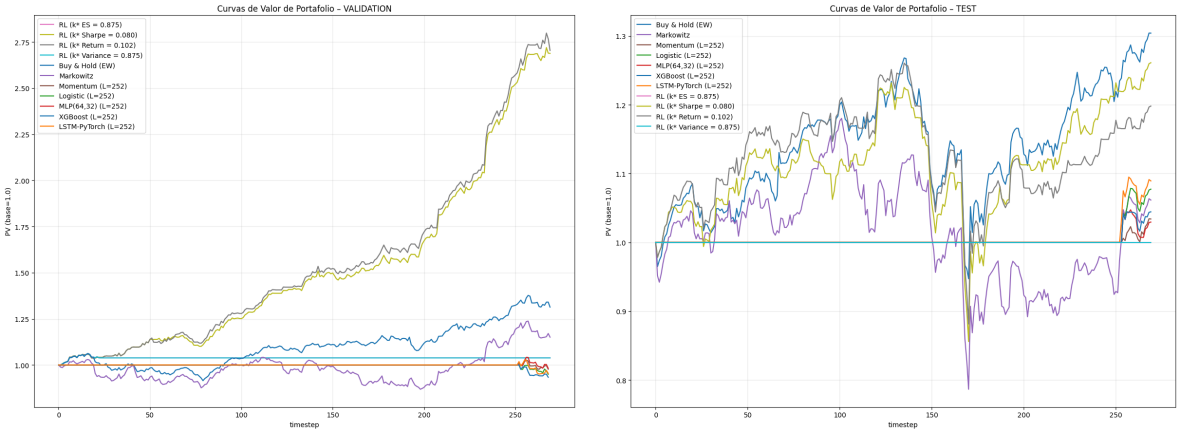
Si el objetivo es maximizar retorno en test con buen ajuste riesgo–retorno, **RL–Sharpe** y **RL–Return** son atractivas; si la prioridad es *protección de cola*, **Momentum** domina (menor ES_{95}) acorde a la iteración anterior.

La Figura 6.5 y la Tabla 6.5 muestran que para **AAPL–JPM** con $L = 20$, **RL–Sharpe** ofrece el mejor Sharpe fuera de muestra, mientras que **RL–Return** y **Buy&Hold** compiten en retorno absoluto. Por otro lado, **Momentum** reduce mejor el riesgo de cola

y se mantiene estable en el cambio de régimen. Las políticas **RL–ES/Var** requieren recalibración para no perder tendencia en test, ya que su alta prudencia limita la exposición y las oportunidades de captura de tendencia.

Tercera iteración sobre las acciones

La iteración final sobre acciones se presenta en la Figura 6.6, que ilustra la comparación de curvas de valor de portafolio en validación y testeo.



(a) Validación *(b) Testeo*

Figura 6.6. Comparación de curvas de valor de portafolio para la iteración 6.

Estrategia	VALIDACIÓN				TESTEO			
	Ret acum	Sharpe	Var	ES95	Ret acum	Sharpe	Var	ES95
RL (k* ES = 0.875)	0.037	1.037	0.000	0.000	0.000	0.000	0.000	0.000
RL (k* Variance = 0.875)	0.037	1.037	0.000	0.000	0.000	0.000	0.000	0.000
MLP(64,32) (L=252)	-0.018	-0.328	0.000	0.000	0.029	0.552	0.000	0.000
Momentum (L=252)	-0.023	-0.454	0.000	0.000	0.033	1.167	0.000	0.000
XGBoost (L=252)	-0.069	-1.390	0.000	0.000	0.043	0.817	0.000	0.000
Logistic (L=252)	-0.052	-1.000	0.000	0.000	0.074	1.327	0.000	0.000
LSTM-PyTorch (L=252)	-0.047	-0.730	0.000	0.000	0.086	1.321	0.000	0.000
RL (k* Sharpe = 0.080)	0.989	6.092	0.000	0.012	0.232	0.840	0.000	0.038
RL (k* Return = 0.102)	0.995	5.943	0.000	0.014	0.181	0.653	0.000	0.039
Buy & Hold (EW)	0.273	1.784	0.000	0.020	0.266	1.015	0.000	0.036
Markowitz	0.140	0.593	0.000	0.030	0.060	0.178	0.000	0.047

Tabla 6.6. Iteración 6: métricas por estrategia.

En **validación**, según la Figura 6.6, las políticas de *RL* calibradas por *Sharpe* y *Return* exhiben una trayectoria casi monótona y consistentemente superior al resto. En

cambio, *RL–ES/Var* permanecen prácticamente planas (exposición nula o casi nula), lo que refleja un umbral de decisión excesivamente conservador, tal como se observó en la iteración anterior. En **testeo**, el comportamiento cambia de forma marcada pues *Buy&Hold* progresa de manera sostenida y *Markowitz* capta parte de la tendencia, mientras que *RL–ES/Var* se mantienen planas (sin operar) y *RL–Sharpe/Return* aportan ganancias moderadas. Los modelos supervisados (*Logistic* y *LSTM*) exhiben curvas suaves con oscilación contenida, fiel a un comportamiento menos errático.

Los resultados de la Tabla 6.6 confirman que en **validación**, *RL–Sharpe* y *RL–Return* dominan claramente (ret. acum. ≈ 0.989 y 0.995; Sharpe ≈ 6.09 y 5.94; $\text{ES}_{95} \approx 0.012$–$0.014$). *RL–ES/Var* apenas avanzan (ret. acum. ≈ 0.037; Sharpe ≈ 1.04), consistente con su línea horizontal en la Figura 6.6. En **testeo**, el patrón se mantiene: **Buy&Hold** alcanza el mayor retorno acumulado (≈ 0.266) y un Sharpe ≈ 1.015, mientras que *RL–Sharpe/Return* logran retornos positivos más moderados (ret. acum. ≈ 0.232 y 0.181; Sharpe ≈ 0.84 y 0.65; $\text{ES}_{95} \approx 0.038$–$0.039$). Entre los modelos supervisados, *Logistic* y *LSTM* destacan con los mejores Sharpes (≈ 1.33 y 1.32) aunque con retornos más contenidos (≈ 0.074–0.086). *Momentum* mantiene un buen balance riesgo–retorno (Sharpe ≈ 1.17; ret. acum. ≈ 0.033). Finalmente, *RL–ES/Var* se mantienen *inactivos* (ret. acum. $= 0.000$).

Respecto a la iteración anterior, la **brecha validación–test** es grande para *RL–Sharpe/Return*: pasan de Sharpe ≈ 6 en validación a ≈ 0.65–0.84 en test, con colas más pesadas ($\text{ES}_{95} \approx 0.038$–$0.039$). Esto sugiere posiblemente *subreactividad* por la ventana anual ($L = 252$). El caso más extremo es *RL–ES/Var*, que en test permanece prácticamente sin exposición (línea horizontal), por lo que no captura la tendencia del periodo, dado principalmente por el alto umbral τ.

Siguiendo con la idea anterior y observando tanto la Figura 6.6 como la Tabla 6.6, con $L = 252$ las señales se vuelven **lentas**; políticas con umbrales estrictos (p. ej., por ES o Var) pueden terminar *no operando* en test. Por lo que, si el objetivo es **retorno fuera de muestra**, *Buy&Hold* y *RL Sharpe* domina en este tramo; si se prioriza **eficiencia riesgo–retorno**, *Logistic*, *LSTM* y *Momentum* logran los mejores Sharpes, aunque sacrifican retorno absoluto.

Con $L = 252$, las políticas *RL–Sharpe/Return* continúan siendo efectivas en validación y mantienen desempeño aceptable en test. En contraste, *RL–ES/Var* permanecen inactivas por sus altos umbrales, mientras que *Buy&Hold* y *RL–Sharpe/Return* lideran el retorno fuera de muestra. Los modelos *Logistic*, *LSTM* y *Momentum* muestran los

mejores Sharpes, aunque con ganancias más acotadas, posiblemente por la inactividad inicial de 252 pasos y la tendencia positiva del mercado. Así, la evidencia refuerza el *trade-off* entre políticas de crecimiento (RL) y estabilidad de cola (Momentum/LSTM) bajo ventanas largas. La política elegida debe reflejar la preferencia entre *crecimiento* y *estabilidad de cola* bajo ventanas largas, en donde RL no se comporta de tan buena forma en comparativa con ventanas más cortas.

El análisis de las iteraciones 4, 5 y 6, que evalúa el mismo periodo que las criptomonedas, revela un **comportamiento más estable y predecible** en el mercado de acciones. A diferencia de la alta volatilidad del mercado cripto, las políticas del agente de RL mostraron una mejor capacidad de generalización. Aunque las estrategias agresivas de RL (optimizadas para Sharpe y retorno) también dominaron la fase de validación, su desempeño en la fase de testeo se mantuvo competitivo, no experimentando la caída brusca observada con las criptomonedas.

La brecha de rendimiento entre la validación y el testeo fue menos pronunciada, lo que sugiere una mayor **robustez** de los modelos de RL en un entorno con menor riesgo de cola. Esto se evidenció con la competitividad de *Buy & Hold* y la *LSTM*, que lograron retornos sólidos y un buen control de riesgo. En este mercado, las estrategias de *machine learning* y RL demostraron una ventaja competitiva, obteniendo retornos superiores sin comprometer la gestión de riesgo, lo que contrasta con el perfil de las criptomonedas, donde el ***trade-off*** **entre rentabilidad y riesgo** es mucho más acentuado.

6.4. Resultados sobre Rentabilidad: ETF

(a) Validación *(b) Testeo*

Figura 6.7. Comparación de curvas de valor de portafolio para la iteración 7.

Estrategia	VALIDACIÓN				TESTEO			
	Ret acum	Sharpe	Var	ES95	Ret acum	Sharpe	Var	ES95
RL (k* ES = 0.226)	0.537	3.899	0.000	0.011	0.074	0.483	0.000	0.017
RL (k* Variance = 0.226)	0.537	3.899	0.000	0.011	0.074	0.483	0.000	0.017
RL (k* Sharpe = 0.111)	0.689	4.730	0.000	0.012	0.067	0.320	0.000	0.020
RL (k* Return = 0.111)	0.689	4.730	0.000	0.012	0.067	0.320	0.000	0.020
Momentum (L=10)	0.062	0.465	0.000	0.014	-0.000	-0.003	0.000	0.018
Buy & Hold (EW)	0.177	1.204	0.000	0.014	0.119	0.727	0.000	0.017
XGBoost (L=10)	0.410	2.660	0.000	0.015	0.007	0.043	0.000	0.018
Markowitz (max μ, TRAIN)	0.364	2.227	0.000	0.015	0.232	0.927	0.000	0.027
LSTM-PyTorch (L=10)	0.170	1.127	0.000	0.015	0.099	0.512	0.000	0.023
MLP(64,32) (L=10)	0.424	2.694	0.000	0.016	-0.055	-0.320	0.000	0.019
Logistic (L=10)	0.072	0.473	0.000	0.016	-0.032	-0.149	0.000	0.021

Tabla 6.7. Iteración 7: métricas por estrategia.

En **validación**, tal como se observa en la Figura 6.7, las variantes de RL calibradas para maximizar *Sharpe/Return* ($k^* = 0.111$) presentan trayectorias crecientes y dominan ampliamente al resto; *RL–ES/Var* ($k^* = 0.226$) muestran curvas más parsimoniosas, con ganancia sostenida y menor oscilación, evidenciando un comportamiento defensivo frente a cambios de régimen. En **testeo**, tras el shock común ($t \approx 250$), se reordena el ranking: *Markowitz* lidera el rebote y cierra con el mayor nivel de PV; *Buy&Hold* y *LSTM* acompañan con recuperación estable, mientras que las varian-

tes de RL quedan por detrás en crecimiento, manteniendo, eso sí, oscilaciones más contenidas cuando se optimiza por ES/Var.

A partir de la Tabla 6.7, se observa que en validación *RL–Sharpe/Return* logra el mejor desempeño agregado (ret. acum. ≈ 0.689, Sharpe ≈ 4.730) con colas controladas ($ES_{95} \approx 0.012$). Por su parte, *RL–ES/Var* obtiene un retorno acumulado sólido (≈ 0.537) con la *mejor protección de cola* ($ES_{95} \approx 0.011$) y la menor varianza, coherente con la naturaleza conservadora de su criterio. Los demás benchmarks, como *XGBoost* y *MLP*, se ubican en un punto intermedio, mientras que *Buy&Hold*, *LSTM*, *Momentum* y *Logistic* quedan claramente por debajo en casi todas las métricas.

En **Testeo**, según la Figura 6.7 y la Tabla 6.7, *Markowitz* obtiene el mayor índice de Sharpe (≈ 0.927) y el mayor retorno acumulado (≈ 0.232), aunque con la peor cola ($ES_{95} \approx 0.027$) y la varianza más alta. Por otro lado, *Buy&Hold* y *LSTM* muestran mayor equilibrio, pues los retornos acumulados ≈ 0.119 y 0.099 y Sharpe ≈ 0.728 y 0.512 se mantienen de mejor forma que los demás benchmarks. Los *RL–ES/Var* preservan el **perfil defensivo** de su naturaleza, el $ES_{95} \approx 0.017$ se transforma en la mejor cola junto a *Buy&Hold*, y varianzas más bajas del set, pero con crecimiento acotado dado su retorno acumulado de ≈ 0.074. Los *RL–Sharpe/Return* caen en el ranking fuera de muestra, pues su retorno acumulado es ≈ 0.067 y $ES_{95} \approx 0.020$. Finalmente, los modelos clásicos supervisados (*XGBoost*, *MLP*, *Logistic*) evidencian debilidad en test, inclusive dos de ellos con retorno acumulado negativo, siendo un pésimo indicativo.

Las políticas **agresivas** (*RL–Sharpe/Return*) exhiben la mayor brecha entre fases (de retorno acumulado ≈ 0.689 a 0.067), lo que refleja *alta sensibilidad al régimen* fuera de muestra. Por otro lado, las políticas **defensivas** (*RL–ES/Var*) mantienen el carácter de baja varianza y buena cola en test, aunque sacrifican crecimiento. La estabilidad de *Buy&Hold* y *LSTM*, junto con el liderazgo de *Markowitz* tras el shock, sugiere que en este período las estrategias menos dependientes del umbral optimizado en validación (y con mayor flexibilidad de ajuste) generalizan mejor en un entorno de mercado más maduro como el de los ETFs.

Segunda iteración sobre ETF

Los resultados para la 2da iteración sobre los ETF (8va iteración total) son:

(a) Validación *(b) Testeo*

Figura 6.8. Comparación de curvas de valor de portafolio para la iteración 8.

Estrategia	VALIDACIÓN				TESTEO			
	Ret acum	Sharpe	Var	ES95	Ret acum	Sharpe	Var	ES95
RL (k* ES = 0.173)	0.443	4.016	0.000	0.013	0.081	0.503	0.000	0.023
RL (k* Variance = 0.357)	0.273	2.547	0.000	0.013	0.082	1.137	0.000	0.010
Momentum (L=20)	0.083	0.816	0.000	0.014	-0.077	-0.643	0.000	0.019
Logistic (L=20)	0.076	0.654	0.000	0.015	-0.114	-0.783	0.000	0.024
Markowitz	0.227	1.849	0.000	0.015	0.203	0.975	0.000	0.029
Buy & Hold (EW)	0.121	1.003	0.000	0.015	0.069	0.528	0.000	0.018
LSTM-PyTorch (L=20)	0.150	1.274	0.000	0.016	0.060	0.390	0.000	0.023
RL (k* Return = 0.010)	0.519	4.429	0.000	0.016	-0.055	-0.422	0.000	0.019
RL (k* Sharpe = 0.010)	0.519	4.429	0.000	0.016	-0.055	-0.422	0.000	0.019
MLP(64,32) (L=20)	-0.017	-0.152	0.000	0.016	0.049	0.389	0.000	0.017
XGBoost (L=20)	0.038	0.310	0.000	0.016	0.062	0.415	0.000	0.016

Tabla 6.8. Iteración 8: métricas por estrategia.

En **validación**, tal como se aprecia en la Figura 6.8, las políticas *RL* muestran fuerte tendencia alcista. Pues *RL–Return/Sharpe* y *RL–ES* crecen de forma sostenida y superan holgadamente a los benchmarks, propio de su naturaleza de seteo. *RL–Var* también sube, pero con pendiente menor, pues en el gráfico se observa una curva más "escalonada", consistente con menor rotación y/o volatilidad medida por desviación estándar. En **testeo** aparece un tramo de estrés a mitad del período y posterior recuperación, *Markowitz* es la curva con avance más persistente y cierra en el nivel de PV más alto, mientras que *RL–Var* y *RL–ES* son quienes aparecen en el podio por detrás del benchmark de Markowitz. Estas estrategias de RL mantienen perfiles más contenidos y estables, pues por ejemplo *RL-Var* a pesar de ser un perfil

más conservador, mantiene un retorno acumulado muy por encima de otras estrategias más agresivas, manteniendo también un riesgo de cola muy bajo, lo que le hace ser la mejor estrategia en términos de riesgo-retorno según el índice de Sharpe. Por otro lado, *RL–Return/Sharpe* pierden tracción y terminan por debajo del desempeño medio, siendo opciones más descartables.

También en **Validación**, según la Tabla 6.8, se puede desprender que las estrategias de *RL–Return/Sharpe* registran los mejores agregados (ret. acum. ≈ 0.519, Sharpe ≈ 4.429, $\text{ES}_{95} \approx 0.016$). *RL–ES* alcanza Sharpe ≈ 4.016 con la mejor cola del grupo ($\text{ES}_{95} \approx 0.013$). Mientras que *RL–Var* queda intermedio (Sharpe ≈ 2.547). El resto de estrategias queda por detrás. En **Testeo**, conforme se aprecia en la Figura 6.8 y la Tabla 6.8, *Markowitz* lidera el *retorno acumulado* (≈ 0.203) con Sharpe ≈ 0.975, pero la estrategia de *RL–Var* ofrece el *mejor Sharpe y mejor ES* fuera de muestra (Sharpe ≈ 1.137, $\text{ES}_{95} \approx 0.010$) aunque con menor retorno (≈ 0.082). La estrategia de riesgo de cola *RL–ES* es competitivo en retorno (≈ 0.081) con riesgo moderado, posiblemente debido a tener un umbral más bajo respecto a la estrategia de RL con varianza, junto con el shock generado entre las observaciones 150 y 200 (la cual la estrategia *RL-ES* logró disminuir su impacto, pero igualmente afectó). En contraste, *RL–Return/Sharpe* muestran desempeño negativo en test (ret. acum. ≈ -0.055), lo que sugiere sensibilidad a cambios de régimen. Entre los benchmarks, *Buy&Hold* y *LSTM* son correctos, mientras que *Momentum* y *Logistic* resultan débiles en este período.

La **brecha validación–test** es mayor en las políticas más agresivas (*RL–Return/Sharpe*), que pasan de Sharpe ≈ 4.4 a valores negativos y aumentan su ES_{95}. Las variantes *RL–ES/Var* preservan mejor su carácter defensivo: *RL–Var* destaca por su control de cola ($\text{ES}_{95} \approx 0.010$) y mejor Sharpe en test, coherente con la curva más parsimoniosa. *Markowitz* captura mejor el rebote y lidera el retorno en test, a costa de una cola más pesada que *RL–Var*.

Bajo los resultados presentados en la Figura 6.8 y la Tabla 6.8, las **estrategias defensivas** como *RL–Var/RL–ES* ofrecen mejor *estabilidad de cola* tanto en validación como en testeo. En particular, *RL–Var* logra el mejor Sharpe y el menor ES_{95}. Si se privilegia retorno, *Markowitz* es la opción. Si se prioriza una estragia más equilibrada o de eficiencia riesgo–retorno, *RL–Var* es preferible que *RL–ES*.

Tercera iteración sobre ETF

Y la iteración final:

(a) Validación *(b) Testeo*

Figura 6.9. Comparación de curvas de valor de portafolio para la iteración 9.

Estrategia	VALIDACIÓN				TESTEO			
	Ret acum	Sharpe	Var	ES95	Ret acum	Sharpe	Var	ES95
Logistic (L=252)	-0.028	-0.880	0.000	0.000	0.011	0.479	0.000	0.000
MLP(64,32) (L=252)	-0.035	-1.030	0.000	0.000	0.016	0.738	0.000	0.000
XGBoost (L=252)	-0.035	-0.926	0.000	0.000	0.013	0.647	0.000	0.000
LSTM-PyTorch (L=252)	-0.018	-0.415	0.000	0.000	0.019	0.757	0.000	0.000
RL (k* Return = 0.088)	0.660	6.384	0.000	0.008	-0.010	-0.062	0.000	0.020
RL (k* ES = 0.088)	0.660	6.384	0.000	0.008	-0.010	-0.062	0.000	0.020
RL (k* Sharpe = 0.088)	0.660	6.384	0.000	0.008	-0.010	-0.062	0.000	0.020
Momentum (L=252)	0.083	0.816	0.000	0.014	-0.077	-0.643	0.000	0.019
RL (k* Variance = 0.391)	0.246	2.419	0.000	0.014	-0.093	-1.279	0.000	0.012
Markowitz	0.227	1.849	0.000	0.015	0.203	0.975	0.000	0.029
Buy & Hold (EW)	0.121	1.003	0.000	0.015	0.069	0.528	0.000	0.018

Tabla 6.9. Iteración 9: métricas por estrategia.

En **validación**, tal como se aprecia en la Figura 6.9, las políticas de *RL* optimizadas por *Return/Sharpe/ES* muestran una trayectoria casi monótona y muy superior al resto (crecimiento estable y baja variabilidad), siendo una fiel imagen de lo que fue para las otras 8 iteraciones anteriores, mientras que *RL–Var* también sube pero con pendiente menor. En **testeo**, el comportamiento cambia y *Markowitz* nuevamente captura mejor la tendencia alcista del periodo y es la curva que más avanza. *Buy&Hold* también progresa, pero en una menor escala. Las variantes *RL–Return/Sharpe/ES*

permanecen prácticamente *planas* gran parte del tramo (exposición muy persistente/casi nula), y *RL–Var* evidencia un perfil escalonado descendente, consistente con señales lentas por la ventana larga.

También, numéricamente en **Validación**, según la Tabla 6.9, *RL–Return/ES/Sharpe* empatan como las mejores (ret. acum. ≈ 0.660, Sharpe ≈ 6.384, $\text{ES}_{95} \approx 0.008$), dado que poseen el mismo umbral de transacción (comportamiento del RL anómalo que no se había repetido anteriormente), lo cual puede sugerir que implementar una ventana de 252 días podría ser innecesario y estaría llenando al agente de información no necesaria que afecte de manera negativa a sus propias políticas. Siguiendo con RL, *RL–Var* queda detrás en retorno (ret. acum. 0.246, Sharpe 2.419). Entre benchmarks, *Markowitz* y *Buy&Hold* aportan retornos moderados. Mientras que las estrategias con modelamiento supervisado, permanecen sin acción durante el tiempo de la ventana, y para cuando comienzan a transaccionar, su rendimiento no es óptimo, concluyendo con resultados negativos para tanto la MLP, Logistic, XGBoost y LSTM. En **Testeo**, conforme a lo observado en la Figura 6.9 y la Tabla 6.9, el ranking se reorganiza de la misma forma que ha sido para las últimas 2 iteraciones, pues *Markowitz* lidera en retorno acumulado (≈ 0.203) con Sharpe ≈ 0.975, *Buy&Hold* también es positivo. *LSTM* y *MLP* presentan retornos pequeños pero con Sharpe razonable (≈ 0.757 y 0.738). En cambio, *RL–Return/ES/Sharpe* quedan *casi neutras o levemente negativas* (ret. acum. ≈ -0.010), y *RL–Var* cae a ret. acum. ≈ -0.093 con Sharpe negativo. Reafirmando de que el agente evaluado y desarrollado en esta tesis, funciona de mejor forma para ventanas más cortas.

Tal como se confirma en la Figura 6.9 y en la Tabla 6.9, y por la propia naturaleza de la metodología de esta tesis, la brecha validación–test es **grande** para las variantes agresivas de *RL* (Return/Sharpe/ES), pues pasan de Sharpe ≈ 6.4 a valores ligeramente negativos, lo que sugiere un posible sobreajuste de validación y/o *subreactividad* de las señales con $L = 252$, con menores turnover o decisiones tardías. *RL–Var*, pese a su buen desempeño en validación, sufre en test, pues por el contrario, *Markowitz* generaliza mejor y *Buy&Hold* funciona como ancla simple al riesgo de mercado en este periodo.

No es difícil de inferir que la **Ventana larga (**$L = 252$**)** reduce ruido pero también *ralentiza* las señales para el modelo de RL, pues las políticas de *RL* con reglas conservadoras pueden pasar gran parte del test *sin ajustar exposición*, perdiendo la tendencia del tramo. Por ello, a fin de encontrar una estrategia óptima de inversión, si

se busca *retorno*, *Markowitz* domina, si se prioriza *estabilidad*, *Buy&Hold* es una base razonable, dado que su ***ES*** es meucho menor en proporción al de Markowitz. Las variantes *RL* requerirían recalibración u otro enfoque (por ejemplo, menor L o mayor penalización) para mejorar su sensibilidad ante ventanas más largas.

Dadas las iteraciones 7, 8 y 9, se tiene un perfil de riesgo y retorno notablemente distinto a los mercados de criptomonedas y acciones. A lo largo del periodo, el entorno se caracterizó por una **menor volatilidad y una tendencia de crecimiento más estable**. En este contexto, la estrategia ***Buy & Hold*** demostró ser el *benchmark* más robusto, consolidándose como una opción altamente competitiva y eficiente en términos de simplicidad y rendimiento ajustado por riesgo.

Las políticas del agente de RL y los modelos supervisados lograron generar retornos positivos en la fase de testeo, pero su desempeño no justificó la complejidad algorítmica y computacional. La principal conclusión en este segmento de activos es que, en mercados con baja volatilidad y fuerte tendencia, las estrategias pasivas o de reglas simples como ***Buy & Hold*** y ***Markowitz*** son a menudo las más efectivas. Esto evidencia que la sofisticación de un modelo debe ser proporcional a la complejidad del entorno de inversión, y en el caso de los *ETFs*, una aproximación más conservadora puede resultar la más óptima.

6.5. Inferencia y robustez estadística

Los valores puntuales de retorno, Sharpe y ES son informativos pero, en muestras con dependencia temporal y colas pesadas, deben acompañarse de medidas de *incertidumbre*. Para ello, seguimos el esquema descrito en el capítulo 2: Primeramente construimos **intervalos de confianza (IC) del Sharpe** mediante *bootstrap por bloques móviles* (MBB), y en segundo lugar se contrasta el **exceso medio de retorno** frente a Buy&Hold (B&H) usando errores **HAC** (Newey–West). Además, se reporta **rotación anualizada** (*turnover*) y **% de días expuesto** para capturar parsimonia operativa.

En esta sección se aplica el protocolo a un caso representativo: **Crypto (BTC–ETH),** L=10 **en *test***. Los IC del 95 % del Sharpe se obtienen con MBB (percentil simple) usando B=1000 réplicas y longitud de bloque $\ell = \lceil T^{1/3} \rceil$; se verifica la estabilidad al variar $\ell \in [5, 20]$. El contraste HAC del exceso medio se implementa como una regresión con constante $d_t = R_t^{(\mathrm{estr})} - R_t^{(\mathrm{B\&H})}$, empleando kernel de Bartlett y $q = 5$

rezagos (también se probó $q \in [3, 10]$ sin cambios sustantivos). El "Exceso vs B&H" está expresado en *retorno log por periodo* de la serie evaluada. Por reproducibilidad, se fija semilla pseudoaleatoria y se registra (B, ℓ, q) por corrida.[1]

Estrategia	Sharpe	IC95 lo	IC95 hi	Exceso vs B&H	t_{HAC}	p_{HAC}	lags	Turnover	% días exp.
RL (k^* ES $= 0.18$)	0.2418	-0.0074	0.6008	0.0021	1.7032	0.0901	5	0.5122	0.8409
RL (k^* Var $= 0.145$)	0.2418	-0.0374	0.5259	0.0012	1.6122	0.0969	5	0.5499	0.8622
RL (k^* Return $= 0.037$)	0.0118	-0.2575	0.2897	-0.0012	-1.4630	0.1435	5	3.5550	0.7474
RL (k^* Sharpe $= 0.012$)	0.0188	-0.2750	0.3359	-0.0011	-1.2495	0.2115	5	4.1944	0.7398
Logistic (L=10)	0.2143	-0.0952	0.5149	0.0005	0.6883	0.4913	5	4.0665	0.8265
MLP(64,32) (L=10)	0.1900	-0.1060	0.4807	0.0005	0.5838	0.5593	5	5.1918	0.7143
Markowitz	0.1911	-0.0622	0.4902	0.0002	0.4120	0.6803	5	0.0000	1.0000
LSTM-PyTorch (L=10)	0.1687	-0.1151	0.4452	0.0002	0.3733	0.7089	5	2.5831	0.9745
Momentum (L=10)	0.1951	-0.1089	0.5010	0.0002	0.2231	0.8235	5	1.5217	0.6735
XGBoost (L=10)	0.1799	-0.1243	0.4980	0.0002	0.2143	0.8303	5	4.7442	0.7474
Buy & Hold (EW)	0.1410	-0.1185	0.4460	0.0000	—	—	5	0.0000	1.0000

Tabla 6.10. Crypto (BTC–ETH), L=10 (Test). Otra iteración.

Los IC del Sharpe son amplios y, en esta iteración, *incluyen cero* para todas las estrategias por sí solos no acreditan Sharpe significativamente positivo al 5 %, sin embargo la estrategia de RL mediante la minimización del ES_α se muestra como la mejor en términos de exceso medio de retorno y en intervalo de confianza más "positivo".

Ninguna estrategia muestra un exceso medio frente a B&H estadísticamente distinto de cero al 5 % ($p_{HAC} > 0.05$), aunque las variantes **RL–ES/Var** quedan más cercanas al umbral ($p \approx 0.09$) y con el mayor Sharpe puntual (0.242). *(iii)* Las configuraciones **agresivas** (RL–Return/Sharpe) presentan exceso medio negativo (no significativo) y mayor *turnover* (3.6–4.2), consistente con whipsaws. *(iv)* **Markowitz/Momentum/LSTM** se ubican en un punto intermedio: Sharpes puntuales en 0.17–0.24, p_{HAC} altos y perfiles de parsimonia acordes a su diseño (p. ej., Markowitz con 100 % de exposición y rotación nula).

Se aplica exactamente el mismo protocolo en Equities ($L \in \{10, 20\}$), ETFs (L=20) y Crypto ($L \in \{20, 252\}$). Los *patrones cualitativos* se repiten, donde **RL–Var/ES** tienden a exhibir IC más concentrados y menor rotación. Las configuraciones **agresivas** muestran IC más anchos y, a menudo, exceso medio no significativo frente a B&H.

[1] Nótese que, dado que se comparan múltiples estrategias, los p-valores HAC son *marginales*.

Finalmente, la **parsimonia operativa** se asocia a mayor estabilidad fuera de muestra. Para no sobrecargar el cuerpo del texto, se deja este caso representativo.

6.6. Umbrales dinámicos y actividad de trading

Se vincula la metodología con el comportamiento observado inspeccionando el *umbral de activación* dinámico,

$$\tau_t \;=\; k\,\hat{\sigma}_t,$$

y los *heatmaps* de posiciones. Las gráficas de las figuras 6.10, 6.11, 6.12, 6.13 utilizan *Cripto* con $L{=}10$ como caso ilustrativo; los patrones se generalizan a los pares de ETF y acciones.

Un L más largo o un k mayor inflan τ_t, amortiguan la intensidad de trading y retrasan la adaptación de régimen, mientras que un L más corto / k menor aumentan la respuesta al costo de mayor riesgo de *whipsaw* o cambio rápido de precio. BTC suele presentar $\hat{\sigma}_t$ mayor que ETH, lo que eleva su umbral por activo.

Los paneles 6.10 y 6.11 muestran que τ_t escala tanto con k como con la estimación de volatilidad $\hat{\sigma}_t$ (EWMA/rodante). Tras choques de volatilidad, $\hat{\sigma}_t$ sube rápido y decae lento (histeresis), manteniendo τ_t elevado y suprimiendo entradas incluso cuando $|r_t|$ se normaliza. Por eso, configuraciones conservadoras (alto k, L largo) producen segmentos tipo meseta y menor *turnover* [2] mientras que configuraciones agresivas (bajo k, L corto) reingresan a tendencias más rápido pero pagan más *whipsaws*, como en Cripto (Sección 6.2).

Si se observa el panel 6.11, se puede notar cierta Asimetrías entre activos y sesgo inductivo. Pues en BTC la línea discontinua (τ_t) se ubica por encima de la mayoría de $|r_t|$ con k alto, induciendo tramos largos en *flat*; en ETH, su mayor volatilidad de corto plazo puede empujar τ_t aún más, reforzando la desexposición. Los *heatmaps* de 6.12 y 6.13 muestran la microestructura de decisiones: RL–Sharpe presenta más tramos planos (selectividad para preservar homogeneidad de riesgo), mientras que RL–Return mantiene posiciones largas más frecuentes (escala de crecimiento). Estos patrones micro explican los resultados macro: en Cripto con $L{=}20$, RL–Return/ES se mantienen *competitivas en retorno* pero con colas más pesadas, mientras que una regla Momentum lidera en Sharpe y ES_{95} (Tabla 6.2).

[2] en línea con los "escalones" observados para RL–Var en ETFs (Sección 6.4)

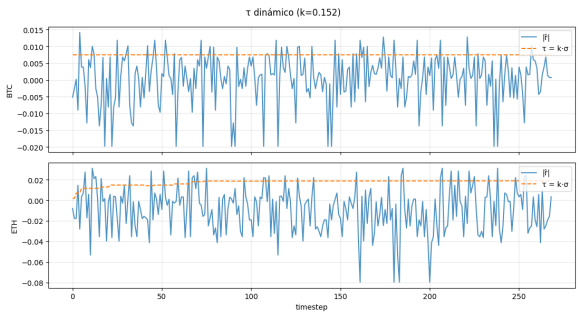


Figura 6.10. k bajo (≈ 0.15): τ_t más estrecho, más operaciones y adaptación de régimen más rápida.

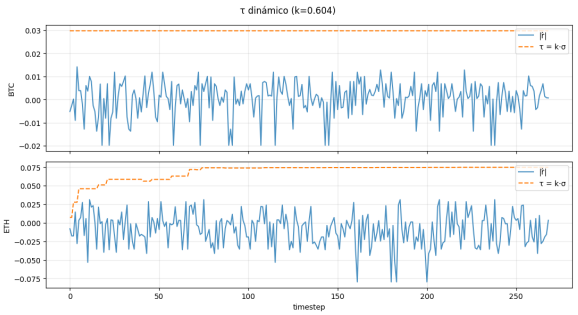


Figura 6.11. k alto (≈ 0.60): τ_t elevado, menos operaciones y riesgo de *under-trading*.

Figura 6.12. Heatmap de posiciones (RL–Sharpe, Test).

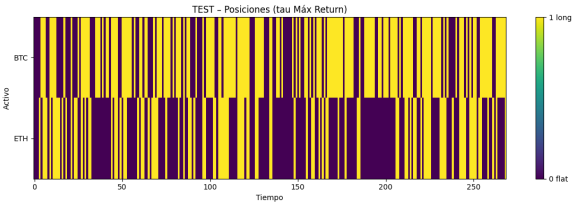


Figura 6.13. Heatmap de posiciones (RL–Return, Test).

Aumentar k debe reducir la densidad de operaciones y las pérdidas de cola (ES_{95}) pero comprimir el retorno acumulado. Mientras que acortar L debe elevar la respuesta y el *turnover* y puede mejorar retornos en fases tendenciales, a costa de mayores *drawdowns*. Para que k no afecte directamente a τ es que se normaliza este mismo τ_t por activo (por ejemplo, fijando una probabilidad objetivo $\mathbf{P}(|r_t| > \tau_t)$) o añadir un proxy de volatilidad/régimen al estado del RL, a fin de reducir el *under-/over-trading*.

Capítulo 7

Conclusiones

A partir cada iteración (3 para cada par de activos), con múltiples ventanas ($L \in \{10, 20, 252\}$) y variantes de la política de RL (umbral k optimizado por Sharpe, Retorno, Varianza y ES), se encuentra un patrón consistente en la mayoría de las iteraciones:

1. En **validación**, el RL con k agresivo (*RL–Return/Sharpe*) domina en retorno y Sharpe. Sin embargo, en **testeo**, su ventaja se reduce.

2. Las variantes **defensivas** (*RL–Var/ES*) son las que **mejor generalizan** en términos de validación y testeo, pues preservan ES y Sharpe razonables en test y el gap no es tan amplio.

3. La estrategia de **Markowitz** tiende a liderar el *retorno absoluto* en test (especialmente en ETF y algunas corridas en cripto), mientras **Momentum/LSTM** frecuentemente logran los *mejores Sharpe* y *colas más livianas*. Se debe entender que, al no imponer una condición respecto al riesgo (obedeciendo a la propia definición de Markowitz), sumado con la maximización del retorno, entonces la asingación de pesos será asignar en su totalidad al activo con mayor retorno en entrenamiento. Por lo que, la utilización de limitantes de capitalización ayudan a prevenir que la estrategia de Markowitz se transforme en un Buy&Hold.

4. El **tamaño de ventana** juega un rol muy importante. Pues, ventanas cortas/medias ($L = 10, 20$) benefician la reactividad del RL, ya que con $L = 252$ aparecen *subreactividad* o *inacción* (umbrales altos).

5. La **parsimonia operativa** (menor rotación, menor % de días expuesto) correlaciona con mayor robustez fuera de muestra, lo cual es algo propio de la naturaleza del trading. Pues a menor exposición de posiciones activas, menor es la probabilidad de exponerse a caídas en el retorno.

Lo anterior viene principalmente evidenciado según las iteraciones I_j en donde:

- I_2 : Momentum obtiene el mejor Sharpe (≈ 1.26) y mejor ES_{95} (≈ 0.046), mientras *RL–ES/Var* mantienen retornos competitivos con colas moderadas, *RL–Return/Sharpe* retroceden respecto de su desempeño de validación.
- I_3 : *RL–Var* y Markowitz lideran el *retorno* en test (RL–Var con $SR \approx 1.03$), mientras que **LSTM** y **Momentum** maximizan Sharpe (≈ 1.30 y 1.27) y entregan colas más livianas ($\text{ES}_{95} \approx 0.03$).
- I_4 : *RL–Sharpe/Return* logra el *mayor retorno* en test (≈ 0.48), pero **LSTM** presenta el *mejor Sharpe* (≈ 1.49) con ES_{95} contenido. Por otro lado, **Momentum** muestra la *mejor cola* ($\text{ES}_{95} \approx 0.029$).
- I_5 : *RL–Sharpe* lidera el Sharpe en test (≈ 1.22) con retorno competitivo, *RL–ES/Var* quedan casi inactivos por umbrales altos, lo que evidencia del riesgo de calibraciones que *bloquean* la operativa con ventanas largas.
- I_8 : *RL–Var* ofrece el *mejor Sharpe* en test (≈ 1.14) y el *menor ES* ($\text{ES}_{95} \approx 0.01$), mientras que Markowitz maximiza el *retorno* (≈ 0.20).
- I_9 : Markowitz lidera el retorno (≈ 0.203) y *RL–Return/Sharpe/ES* quedan planos (señal de subreactividad y/o umbralización excesiva con L largo).

Lo anterior está alineado con la intuición de que el *régimen* de mercado entre validación y test no es estable, y que las políticas más *parsimoniosas* y con *control explícito de cola* generalizan mejor, sobretodo dado el horizonte temporal utilizado descrito por distintas fases sobre las cuales transaron los activos.

Los hallazgos obtenidos a partir de los resultados de esta tesis, se conectan con estudios validados de revistas dedicadas a las finanzas cuantitativas. En particular, acerca de la conexión de los hallazgos:

1. Estudios como Jiang et al. (2017) reportan superioridad de DRL en determinados entornos y periodos, pero bajo supuestos y ventanas específicos. Kolm and Ritter (2019) discuten que la no estacionariedad, los cambios de régimen y los costos friccionales erosionan la ventaja en despliegue real. Los resultados de esta tesis replican este *efecto de régimen*, lo que domina en validación (RL agresivo) no siempre traslada al tramo de test.

2. Fischer and Krauss (2018) muestran que LSTM puede superar a enfoques lineales en predicción direccional. En esta tesis, LSTM y Momentum suelen maximizar *Sharpe* y reducir ES en test (especialmente con L más largos), siguiendo que sus reglas implícitas son más estables y menos dependientes de una calibración fina de umbrales.

3. La inferencia sobre el Sharpe es frágil bajo colas pesadas y autocorrelación (Lo, 2002). Por ello se complementa con medidas coherentes como el ES (Artzner et al., 1999) y con filtros de volatilidad EWMA —incluida una versión de dos semividas (*two-speed*)— inspirados en RiskMetrics Group (1996); McNeil et al. (2015). En nuestras pruebas, las variantes *RL–Var/ES*, que incorporan ese control de cola, preservan con mayor estabilidad la relación riesgo–retorno fuera de muestra.

Algunas causas probables para este gap producido entre validación y test, son por ejemplo la no estacionariedad y cambio de régimen. Pues la reordenación sistemática del ranking en test dice que los patrones explotados por k (seleccionado en validación) no persisten. El fenómeno es más agudo con $L = 252$, cuando las señales se vuelven lentas y los umbrales bloquean la operativa.

También, probablemente recompensar sólo al final del episodio con Sharpe anualizado dificulta la asignación de crédito paso a paso (crédito diferido), potenciando sobreajuste a trayectorias de validación. Mientras que optimizar k para maximizar Sharpe/Retorno en un único split puede elegir soluciones frágiles (“picos” en la grilla) con pobre *transferencia* a test. Se observan episodios con *inacción* prolongada (RL–ES/Var con $L = 252$) y, en el extremo opuesto, políticas demasiado activas con mayor ES en test.

7.1. Soluciones para trabajo futuro

Respecto al aprendizaje y las recompensas del modelo, se podrían utilizar algunas mejoras, como por ejemplo:

- Combinar Sharpe con penalización por ES o drawdown *intra-episodio* para guiar el aprendizaje.
- **Distributional/CVaR-RL**: optimizar directamente colas (p. ej., CVaR) para mejorar robustez fuera de muestra en mercados con colas pesadas. O bien aplicar el análisis de EVT mostrado a en el capítulo 2, pero que finalmente no fue aplicado.
- **Acción continua con sizing**: pasar de $\{-1, 0, 1\}$ a pesos continuos con presupuesto de riesgo y penalización por cambio de posiciones en la recompensa.

Respecto a la calibración de parámetros e hiperparámetros, algunas ideas que podrían mejorar el modelo, son:

- **Walk-forward anidado**: recalibrar el mapeo $s \mapsto (\mu, \sigma)$ y k en ventanas deslizantes; evitar elegir k en un único split. Lo cual es ampliamente utilizado en modelos financieros.
- **Regularización de** k: imponer *early stopping* y suavizar la función objetivo (promedio de métricas o restricciones sobre active ratio).

Incluir otros datos de mercado o sus características, a modo de simular un entorno más similar al real. Por ejemplo, mediante la inclusión de *slippage*, spreads y liquidez, a fin de penalizar explícitamente el turnover en la recompensa.

Implicancias prácticas por perfil de inversor

Dados los resultados y la amplia gamma de estrategias probadas, se podría dividir la clasificación según la posición de aversión al riesgo que tenga el inversionita/trader en particular. Algunas divisiones, serían:

- **Crecimiento absoluto (tendencial):** Markowitz y, en ventanas cortas, *RL–Return/Sharpe* (aceptando mayor ES).

- **Eficiencia riesgo–retorno / protección de cola:** *RL–Var/ES*; alternativamente Momentum/LSTM cuando se busca estabilidad de Sharpe.
- **Base estable y sencilla:** *Buy&Hold* como overlay que evita largos periodos en cero exposición cuando el RL queda bloqueado por umbrales.

En **ETF,** $L = 20$ la combinación *RL–Var* + overlay pasivo fue la más cercana al "equilibrio" (Sharpe alto, ES bajo), mientras que en **cripto** la rotación del ranking entre validación y test justifica detectores de régimen y recalibración frecuente; en **acciones** con $L = 10$–20, *RL–Sharpe/Return* muestran buen retorno, pero LSTM/Momentum lideran Sharpe en shocks.

Síntesis final

En esta tesis se realiza una evaluación comparativa amplia, pues la estrategia de RL se comparó con múltiples *benchmarks*, divididos según clásicos y modelos supervisados, mediante la utilización de métricas de **riesgo de cola** (ES), **parsimonia operativa** (turnover, % días expuesto) y **Sharpe** en validación y test. Además, se realiza la identificación de **reglas defensivas** (*RL–Var/ES*) como núcleo para despliegue, y de ventanas cortas/medias como condición dominante de éxito para RL, no necesitando grandes volúmenes de información para tomar mejores decisiones, lo que le permite ser un modelo especialmente mejor en términos de costo computacional, pudiendo ser una buena herramienta para modelos de finanzas o de trading donde se trabaje con datos intradía o con una frecuencia más alta.

Este RL con recompensa en Sharpe y umbrales de riesgo dinámicos es competitivo y económicamente consistente frente a otros *benchmarks*. Su ventaja práctica y eficiencia computacional se muestra mejor en versiones defensivas (*RL–Var/ES*) y con ventanas cortas–medias, donde muestra **mejores colas** y **Sharpe** estables en test. Cuando el objetivo es *retorno absoluto* en tramos tendenciales, **Markowitz** tiende a liderar, mientras que si la prioridad es *eficiencia riesgo–retorno*, **Momentum** o **LSTM** son referentes sólidos. La clave para cerrar la brecha validación–test es **adaptar** k y el umbral τ al *régimen* mediante *walk-forward*, integrando además fricciones desde el diseño.

7.2. Apéndice

Proposición 1: Existe $X, Y \in \mathcal{L}$ y un nivel $\alpha \in (0,1)$ tales que

$$\mathrm{VaR}_\alpha(X+Y) \;>\; \mathrm{VaR}_\alpha(X) + \mathrm{VaR}_\alpha(Y).$$

Demostración: Considérese un espacio de probabilidad con dos eventos disjuntos A, B de probabilidad $p \in (0, \frac{1}{2})$. Defínanse las pérdidas

$$X = \mathbf{1}_A, \qquad Y = \mathbf{1}_B.$$

Así, X (resp. Y) toma el valor 1 con probabilidad p y 0 con probabilidad $1-p$. Para cualquier nivel α tal que $1-p \geq \alpha > 1-2p$, se tiene:

$$\mathbf{P}(X \leq 0) = 1-p \geq \alpha \quad \Rightarrow \quad \mathrm{VaR}_\alpha(X) = 0,$$

y análogamente $\mathrm{VaR}_\alpha(Y) = 0$. Sin embargo, como A y B son disjuntos, $X+Y$ vale 1 en $A \cup B$ (probabilidad $2p$) y 0 en el complemento (probabilidad $1-2p$). Dado que $1-2p < \alpha$, y $1 = \text{mín}\{c : \mathbf{P}(X+Y \leq c) \geq \alpha\}$, se obtiene que:

$$\mathrm{VaR}_\alpha(X+Y) = 1 \quad > \quad 0+0 = \mathrm{VaR}_\alpha(X) + \mathrm{VaR}_\alpha(Y),$$

lo que viola la subaditividad.

Proposición 2: Sea $R_X, R_Y \in L^1$ y sea $R_Z := R_X + R_Y$. Para $\alpha \in (0,1)$, con $x_{(\alpha)} := VaR_\alpha(X)$, $y_{(\alpha)} := VaR_\alpha(Y)$, $z_{(\alpha)} := VaR_\alpha(Z)$ y $ES_\alpha(X)$ definido anteriormente, entonces la medida de riesgo $\rho(R) := ES_\alpha(R)$ es subaditiva:

$$\rho(Z) \;\leq\; \rho(X) + \rho(Y), \quad \text{i.e.} \quad ES_\alpha(Z) \;\leq\; ES_\alpha(X) + ES_\alpha(Y).$$

Demostración: Por la linealidad de la esperanza, se tiene que:

$$(1-\alpha)[ES_\alpha(X) + ES_\alpha(Y) - ES_\alpha(Z)] = \mathbb{E}\Big[R_Z \mathbf{1}_{\{R_Z \leq z_\alpha\}} - R_X \mathbf{1}_{\{R_X \leq x_\alpha\}} - R_Y \mathbf{1}_{\{R_Y \leq y_\alpha\}}\Big]$$

Luego, usando la definición de que $Z = X + Y$, se tiene que lo anterior es igual a:

$$\mathbb{E}\Big[R_X\big(\mathbf{1}_{\{R_Z \leq z_\alpha\}} - \mathbf{1}_{\{R_X \leq x_\alpha\}}\big) + R_Y\big(\mathbf{1}_{\{R_Z \leq z_\alpha\}} - \mathbf{1}_{\{R_Y \leq y_\alpha\}}\big)\Big].$$

Si $R_X > x_\alpha$, entonces $\mathbf{1}_{\{R_Z \leq z_\alpha\}} - \mathbf{1}_{\{R_X \leq x_\alpha\}} \geq 0$; si $R_X \leq x_\alpha$, la diferencia es ≤ 0. Por lo tanto, se tiene que

$$R_X\big(\mathbf{1}_{\{R_Z \leq z_\alpha\}} - \mathbf{1}_{\{R_X \leq x_\alpha\}}\big) \;\geq\; x_\alpha\big(\mathbf{1}_{\{R_Z \leq z_\alpha\}} - \mathbf{1}_{\{R_X \leq x_\alpha\}}\big),$$

y análogamente para Y. Tomando esperanza y usando por definición que $\mathbf{P}(R_Z \leq z_\alpha) = \mathbf{P}(R_X \leq x_\alpha) = \mathbf{P}(R_Y \leq y_\alpha) = 1 - \alpha$, obtenemos

$$(1-\alpha)[ES_\alpha(Z) - ES_\alpha(X) - ES_\alpha(Y)] \;\geq\; x_\alpha(1-\alpha-(1-\alpha)) + y_\alpha(1-\alpha-(1-\alpha)) = 0.$$

Luego $ES_\alpha(X+Y) \leq ES_\alpha(X) + ES_\alpha(Y)$.

Bibliografía

Acerbi, C. and Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

Bai, Y., Liu, Z., Ma, S., and Li, Y. (2024). A review of reinforcement learning applications in finance: Methods, challenges, and future directions. *arXiv preprint arXiv:2411.12746*.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ.

Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. https://gym.openai.com/. Accessed: 2025-10-23.

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

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

Christoffersen, P. F. (2014). *Elements of Financial Risk Management*. Academic Press, San Diego, 2nd edition.

Christoffersen, P. F. and Diebold, F. X. (2006). Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science*, 52(8):1273–1287.

Chu, J., Chan, S., Nadarajah, S., and Osterrieder, J. (2017). Garch modelling of cryptocurrencies. *Journal of Risk and Financial Management*, 10(4):17.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236.

Cushing, D. L. (2000). Stock returns and trading at the close. *Journal of Financial Markets*, 3(1):45–67.

De March, H. and Lehalle, C.-A. (2018). On the optimal use of trading signals: Trading frequency, speed and bandwidth. *Market Microstructure and Liquidity*, 4(1):1850008.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business*, 38(1):34–105.

Fischer, T. (2018). Reinforcement learning in financial markets – a survey. Technical report, Institute for the World Economy (IfW Kiel).

Fischer, T. and Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669.

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., and Pineau, J. (2018). An introduction to deep reinforcement learning. In *Foundations and Trends in Machine Learning*, volume 11, pages 219–354.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA.

Hall, P., Horowitz, J. L., and Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ.

Heaton, J. B., Polson, N. G., and Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12.

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

Huck, N. (2009). Pairs selection and outranking: An application to the s&p 100 index. *European Journal of Operational Research*, 196(2):819–825.

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

Jiang, Z., Xu, D., and Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*.

Jorion, P. (2006). *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, New York, 3rd edition.

Jorion, P. (2007). *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, New York, 3rd edition.

Kolm, P., Ritter, G., and Tucker, E. (2020). Deep reinforcement learning for asset allocation. *Journal of Financial Data Science*, 2(4):10–30.

Kolm, P. N. and Ritter, G. (2019). Modern perspectives on reinforcement learning in finance. *SSRN*. Overview of RL applications to intertemporal financial decision-making.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241.

Lahiri, S. (1999). Theoretical comparisons of block bootstrap methods. *Annals of Statistics*, 27(1):386–404.

Li, Y. (2019). Reinforcement learning applications in finance. *arXiv preprint arXiv:1906.06218*.

Liu, R. Y. and Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In LePage, R. and Billard, L., editors, *Exploring the Limits of Bootstrap*, pages 225–248. John Wiley & Sons, New York.

Lo, A. W. (2002). The statistics of sharpe ratios. *Financial Analysts Journal*, 58(4):36–52.

Lo, A. W. and MacKinlay, A. C. (1990). An econometric analysis of nonsynchronous trading. *Journal of Econometrics*, 45(1–2):181–211.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419.

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

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, revised edition.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Moody, J. and Saffell, M. (2001). Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4):875–889.

Moody, J., Wu, L., Liao, Y., and Saffell, M. (1998). Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*, 17(5-6):441–470.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 625–632.

Pavlov, I. P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press, London. Traducción de G. V. Anrep.

Politis, D. N. and White, H. (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews*, 23(1):53–70.

Raffin, A., Hill, A., Gleave, A., et al. (2021). Stable-baselines3: Reliable reinforcement learning implementations. In *Journal of Machine Learning Research*. software library.

RiskMetrics Group (1996). *RiskMetrics – Technical Document*. J.P. Morgan, New York, 4th edition.

Rockafellar, R. T. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41.

Schulman, J., Moritz, P., Levine, S., Jordan, M. I., and Abbeel, P. (2016). High-dimensional continuous control using generalized advantage estimation. In *ICLR*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. In *ArXiv preprint arXiv:1707.06347*.

Sharpe, W. F. (1966). Mutual fund performance. *Journal of Business*, 39(1):119–138.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1st edition.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition.

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. The Macmillan Company, New York.

Ticknor, J. L. (2013). A bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications*, 40(14):5501–5506.

Tsay, R. S. (2010). *Analysis of Financial Time Series*. John Wiley & Sons, 3rd edition.

Vittori, E., Di Persio, L., and Ruocco, G. (2020). Option hedging with risk-averse reinforcement learning. *arXiv preprint arXiv:2010.12245*.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.