Universidad de los Andes
Facultad de Ingeniería y Ciencias Aplicadas

CONSTRUCCIÓN DE PORTAFOLIOS DE INVERSIÓN POR MEDIO DE APRENDIZAJE REFORZADO MULTIAGENTE Y APLICACIONES DE LÓGICA DIFUSA

MILENA BONACIC MARTINIC

Tesis para optar al grado de
DOCTORADO EN CIENCIAS DE LA INGENIERÍA

Profesor Guía: Juan Pérez Retamales
Profesor Co-Guía: Héctor López Óspina

Santiago, 1 de julio de 2025

Dedicatoria

Dedicada a mi Mamá

"Donde la generosidad es profunda, la fuerza se vuelve inquebrantable"

Quiero expresar el más profundo agradecimiento a mis queridos hijos, su amor incondicional ha sido el pilar de mi fortaleza y motivación. Cada logro alcanzado en este trabajo no solo es mío, sino también de ustedes. Mi recompensa fue esa anhelada interrupción que no cambiaría por nada, con un ¡mamá!. Anto... por tu paciencia al preguntarme algo y que yo dijera ah?, por comer al lado mío en silencio solo por compartir un rato, Juli... por hablarme y sin quererlo, arrojarme una maravillosa idea, Flori... por compartir el estudio con tu cuadernito y tu infaltable broma, Robi... por todos esos monitos para reparar encima de mi computador por tantas batallas de superhéroes, Nani... por dejarme tantos dibujitos y flores del parque. Espero que esta tesis, que tanto preguntaban cuando terminaría sea un testimonio de mi gratitud hacia ustedes por todo ese tiempo y comprensión que como mis niños cedieron, espero sea una inspiración para seguir persiguiendo sus propios objetivos con pasión, ¡sean cuales sean!

De forma especial, agradecer a mi querido ayudante de múltiples cátedras; Cristóbal Van der Meer, que más allá de ser un ayudante excepcional, has sido un gran compañero y amigo. Tu esfuerzo, dedicación y compromiso con cada ramo que hicimos ha sido invaluable para el éxito de mis clases, los resultados de los estudiantes e incluso para terminar esta tesis.

Por otra parte, un gran agradecimiento a mis alumnos de Estocásticos, Probabilidades, Estadística II y Álgebra&Cálculo, porque cada clase e interacción con ustedes, ha sido un recordatorio de la importancia de la enseñanza y el intercambio de conocimientos en el camino hacia el crecimiento intelectual. Sus preguntas, que a veces me dejaban pensando camino a casa o que más de alguna vez resolví en sueños durmiendo encima del computador, han enriquecido no solo mi comprensión del tema, sino también mi pasión por la enseñanza y la investigación. Han dejado una huella indeleble en mi trayectoria de formación doctoral y espero que estas palabras sean una modesta muestra de mi aprecio por la contribución desconocida por ustedes y en alguna medida espero haberlos apoyado en la construcción de su propio viaje de formación profesional.

Para terminar, expresar un gran agradecimiento a mi profesor guía, Juan Pérez, por su compromiso inquebrantable, incluso cuando yo me perdía un tiempo, ahí estaba su correo en la bandeja de entrada, recordándome que el lugar de esta tesis en mi vida era el primer cuadrante del plano, con lo importante y lo urgente. Es en este contexto que agradezco también su confianza en mi docencia, donde trabajar con Héctor López resultó ser una de las experiencias académicas más enriquecedoras; su apoyo y confianza hicieron de esa etapa una de las más inspiradoras de todo el doctorado y que bueno finalmente terminó.

Extiendo mi agradecimiento al distinguido claustro académico del programa de doctorado, han otorgado un sello inigualable en mi experiencia doctoral y en mi trayectoria como ingeniero.

M.

Resumen

En los mercados financieros, la incertidumbre, la volatilidad y la falta de información no son excepciones, son la norma. Esta tesis se origina en la convicción de que los modelos clásicos de optimización de portafolios resultan insuficientes para enfrentar la complejidad de los mercados. ¿Qué caminos quedan entonces para construir estrategias de inversión robustas y adaptativas?. Se exploran aquí dos grandes aristas: La optimización multiobjetivo apoyada en la entropía y lógica difusa que permite modelar la ambigüedad, apoyada por un análisis multicriterio para la selección de portafolios y el aprendizaje reforzado profundo, en particular multiagente (MARL), donde 4 agentes especializados (RFN, RFI, RVN, RVI) de una AFP aprenden, maximizan sus retornos, eficiencia y minimizan sus movimientos alineados por un coordinador hacia un objetivo global, quien maximiza el retorno, eficiencia conjunta y favorece la diversificación a través de la reducción de las correlaciones y costos de transacción. El diseño no solo enfrenta a los agentes entre sí en la búsqueda de sus objetivos, sino que los incentiva a colaborar por un objetivo global. Se logra un equilibrio dinámico, donde cada uno maximiza su desempeño individual reconociendo que su mejor estrategia depende de las decisiones de sus compañeros, dando lugar a un portafolio donde la eficiencia emergente supera lo posible bajo enfoques puramente competitivos o cooperativos. Por otro lado, a través de los modelos multicriterio se revela el valor de modelar la incertidumbre desde múltiples ángulos. Los métodos basados en lógica difusa con funciones de pertenencia entrópicas, entregan portafolios menos frágiles que los modelos clásicos, capaces de enfrentar escenarios extremos y de responder a los matices de preferencias y restricciones reales.

La evidencia numérica es elocuente, utilizando datos diarios de la plataforma bloomberg en el período julio 2018 a octubre 2024. El modelo MARL, logra un retorno acumulado(RA) de 2.21, reduce el drawdown máximo (DD) a un -14 % y mejora el índice de Sharpe (SH) a 1.39, siendo notablemente superior a modelos de aprendizaje uniagente (RL), con un RA de 2.11, DD de -18 % y un SH de 1.31.El modelo multicriterio entrópico difuso (MED) registra un RA de 2.15, un DD de -17 % y un SH de 1.24. superando en gran medida a los modelos clásicos como Markowitz, con un RA de 1.63, DD de -22 % y SH de 1,07, siendo comparable con modelos RL pero sin superar a MARL. Con base en los resultados, se plantea un modelo híbrido, a partir de una combinación de estrategias en la que los retornos de cada agente del Modelo RL, son vistos como activos de un portafolio a optimizar mediante un MED , los resultados son competitivos al nivel de MARL, con un RA de 2.18, un DD de -15 % y un SH de 1.33.

Se concluye que la rentabilidad, eficiencia y resiliencia del modelo MARL es sustancialmente mejor. La tensión entre competencia y cooperación se convierte en el motor de innovación y eficiencia en un entorno realista, inspirado en la teoría de juegos y la noción de equilibrio de Nash, se demuestra que la inteligencia distribuida, la competencia-cooperación supervisada y la incorporación explícita de la incertidumbre no solo mejoran los resultados, sino que transforman la propia filosofía de la gestión de capital, cuya metodología es extrapolable a diferentes tipos de problemas de optimización con asignación bajo incertidumbre.

Índice general

Glosario de Siglas y Acrónimos

Sigla	**Nombre**	**Descripción breve**
MDP	*Markov Decision Process*	Modelo formal de decisión secuencial con estados, acciones, recompensas y dinámica de transición.
RL	*Reinforcement Learning*	Modelo donde un agente aprende una política para maximizar recompensa acumulada.
DRL	*Deep Reinforcement Learning*	RL con redes profundas como aproximadores para manejar espacios de alta dimensión.
MARL	*Multi-Agent Reinforcement Learning*	RL con múltiples agentes que cooperan o compiten bajo recompensas locales y/o globales.
Q-Learning	—	Algoritmo *off-policy* que estima $Q(s, a)$ mediante actualización temporal-diferencial.
SARSA	—	Algoritmo *on-policy*; actualiza Q con la acción efectivamente ejecutada.
AC	*Actor–Critic*	Familia con actor (política) y crítico (valor/ventaja); base de varios métodos modernos.
A2C	*Advantage Actor–Critic*	Variante síncrona con estimación de ventaja para reducir varianza.
A3C	*Asynchronous A2C*	Múltiples trabajadores asíncronos para exploración estable y más rápida.
DQN	*Deep Q-Network*	Q-Learning con red profunda, *experience replay* y *target network*.
DDQN	*Double DQN*	Reduce la sobreestimación separando selección y evaluación de acciones.
DDPG	*Deep Deterministic Policy Gradient*	Actor–Crítico *off-policy* para espacios de acción continuos; política determinista.

Sigla	Nombre	Descripción breve
TRPO	*Trust Region Policy Optimization*	Optimización de política con restricción KL para pasos estables.
PPO	*Proximal Policy Optimization*	Aproxima TRPO con objetivo *clipped*; buen equilibrio estabilidad–rendimiento.
NAF	*Normalized Advantage Functions*	Q-learning continuo parametrizando ventaja cuadrática para acciones continuas.
Exp. Replay	*Experience Replay*	Memoria de experiencias para romper correlación temporal y reutilizar datos.
MLP	*Multi-Layer Perceptron*	Red densa empleada como aproximador en varias políticas y críticos.
LSTM	*Long Short-Term Memory*	Red recurrente con memoria para dependencias de largo plazo (series de tiempo).
CNN	*Convolutional Neural Network*	Red convolucional para extracción de patrones espaciales o temporales.
AFP	Administradora de Fondos de Pensiones	Contexto institucional del caso de estudio y universo de activos.
RFN	Renta Fija Nacional	Segmento de bonos locales dentro del esquema de cartera.
RFI	Renta Fija Internacional	Segmento de bonos globales dentro del esquema de cartera.
RVN	Renta Variable Nacional	Segmento de acciones locales dentro del esquema de cartera.
RVI	Renta Variable Internacional	Segmento de acciones globales dentro del esquema de cartera.
TOPSIS	*Technique for Order Preference by Similarity to Ideal Solution*	Método multicriterio que prioriza alternativas cercanas al ideal positivo y lejanas del negativo.
AHP	*Analytic Hierarchy Process*	Método multicriterio basado en jerarquías y juicios pareados.
ELECTRE	*ELimination Et Choix Traduisant la REalité*	Familia de métodos multicriterio por sobreclasificación.
PCA	*Principal Components Analysis*	Reducción de dimensión en componentes ortogonales; factoriza variabilidad común.
MED	Modelo Entrópico Difuso	Selección u optimización con lógica difusa y funciones de pertenencia entrópicas.

Sigla	**Nombre**	**Descripción breve**
RA	Retorno Acumulado	Desempeño total del portafolio en el período analizado (factor/múltiplo).
DD	*Max Drawdown*	Máxima caída desde un pico del valor del portafolio; riesgo a la baja.
SH	*Sharpe Ratio*	Retorno excedente por unidad de volatilidad (desempeño ajustado por riesgo).
EW	*Equiweight*	Estrategia de referencia con pesos equiponderados por activo.

Capítulo 1

Introducción

1.1. Antecedentes

La construcción de portafolios eficientes ha sido una de las piedras angulares en la teoría financiera moderna, desde que Harry Markowitz introdujera el concepto de teoría de portafolios y el principio de la frontera eficiente en 1952 [1]. Su enfoque en la diversificación óptima, basado en la maximización del rendimiento ajustado al riesgo, sigue siendo fundamental. En la actualidad, sin embargo, este problema se aborda desde perspectivas más avanzadas y variadas, como optimizaciones matemáticas sujetas a los objetivos de inversión y herramientas de inteligencia artificial entre las que destacan machine learning y en estudios de frontera reinforcement learning . Ambas metodologías buscan proporcionar soluciones importantes dentro de un entorno financiero cada vez más volátil e incierto, aunque lo hacen desde paradigmas analíticos y computacionales diferentes.

El enfoque de optimización está tradicionalmente basado en técnicas matemáticas que utilizan modelos de optimización convexa [2]. Estos modelos buscan minimizar la varianza de los retornos para un nivel dado de rendimiento esperado, o maximizar el rendimiento para un nivel específico de riesgo [3]. Estas técnicas requieren suposiciones específicas sobre las distribuciones de los retornos de los activos y la correlación entre ellos. Si bien son altamente efectivas en mercados estacionarios, muestran limitaciones cuando se enfrentan a mercados no lineales o volátiles, dentro de los cuales las dinámicas de los precios se desvían de las distribuciones normales asumidas [4].

Por otra parte, el aprendizaje por refuerzo (Reinforcement Learning, RL), como rama del aprendizaje automático, ha ganado prominencia en la última década debido a su capacidad para tomar decisiones óptimas en entornos complejos y dinámicos [5]. En este

enfoque, un agente interactúa continuamente con el mercado (su entorno), ajustando sus acciones en función de la recompensa recibida, buscando maximizar una función de utilidad acumulada a lo largo del tiempo [6]. El aprendizaje por refuerzo ofrece un marco flexible que se adapta mejor a mercados en constante cambio y, a diferencia de la optimización tradicional, no requiere suposiciones paramétricas explícitas sobre las distribuciones de retornos [7].

Si bien las técnicas de optimización matemática proporcionan soluciones coherentes dentro del marco teórico, el aprendizaje por refuerzo, permite una mayor capacidad de adaptación y aprendizaje en tiempo real gracias a su capacidad de comprensión del entorno, de tal manera de reflejar comportamientos que en la práctica no se condicen con la teoría. Este punto es fundamental en el análisis de los mercados, particularmente no estacionarios y de alta volatilidad.

El propósito de esta investigación es por una parte extender los modelos actuales de optimización matemática mediante optimización difusa aplicada a portafolios y por otra desarrollar un modelo multiagente basado en aprendizaje reforzado. Para finalizar con un modelo que integra ambas herramientas construyendo un modelo robusto y eficiente para la optimización y selección de portafolios.

Para comenzar, la primera etapa de este trabajo , corresponde a la selección de portafolios mediante técnicas de optimización multiobjetivo con análisis multicriterio.

Basada en la teoría moderna de portafolios (MPT), iniciada por Markowitz en 1952 [1], las técnicas de selección de portaflios consideran un conjunto de herramientas matemáticas que equilibran óptimamente los rendimientos máximos a un nivel aceptable de riesgo. Para esto, el modelo de media- varianza se utiliza como la base, cuya estructura asume que los inversores muestran un comportamiento racional y poseen información completa.

En este contexto, MPT ha dado lugar a diversas líneas de investigación, considerando; medidas de entropía y, modelos de incertidumbre y ambigüedad de la información a través de lógica difusa. Por ejemplo, la utilización de medidas de entropía aborda el problema de la baja diversificación que puede resultar del modelo de media-varianza, al proporcionar una perspectiva más completa sobre la diversificación del portafolio. Mientras tanto, la lógica difusa proporciona un método fundamental para gestionar la incertidumbre y la ambigüedad en la información, considerando rangos aceptables de valores, tales como el rango en el retorno y la varianza de la cartera de inversiones.

En este ámbito, se estudian modelos que combinan medidas clásicas como retorno, varianza y entropía, en un marco de lógica y entropía difusa bajo múltiples criterios. Como concepto general, la entropía difusa mide la vaguedad o ambigüedad de parámetros flexibles

o imprecisos en sistemas de toma de decisiones. Esta medida se puede interpretar como la ganancia de información en entornos difusos, la cual se puede utilizar para comparar la ambigüedad de diferentes sistemas de toma de decisiones [8].

Es importante destacar que el enfoque de entropía difusa no se ha utilizado previamente en la selección de carteras. Además, de acuerdo a esta investigación, los modelos propuestos que incorporan entropía difusa entregan resultados muy superiores a otros modelos de selección de portafolios.

Matemáticamente, si se considera un conjunto difuso A y una función de membresía difusa $\mu_A(x) \in [0, 1]$, definida sobre un conjunto X, la entropía difusa sobre A queda dada por $H_f(A) = -\sum_{x \in X} \mu_A(x) \log_2(\mu_A(x))$. Esta expresión es similar a la Entropía de Shannon que mide la incertidumbre o desorden de una distribución de probabilidad. Sin embargo, en la entropía difusa consideramos el grado de pertenencia dado por $\mu_A(x)$.

En el trabajo de [9] se utiliza un enfoque de entropía difusa en transporte, maximizando la ganancia de información y superando métodos anteriores al calcular matrices de origen-destino. Los prometedores resultados llevan a examinar su potencial en la selección de portafolios, extendiendo esta medida al problema de optimización de portafolios, incorporando el retorno y la varianza tanto en la función objetivo como en las restricciones de los modelos multiobjetivos a través de funciones de pertenencia entrópicas.

Si bien los desarrollos posteriores a la MPT son amplios, la presente tesis se centra en el enfoque base de media-varianza y sus variantes [10], para efectos de modelos multiobjetivos, enfatizando específicamente la entropía difusa.

El modelo de media-varianza es habitualmente conocido y ha demostrado resultados eficientes en la práctica, pero tiene limitaciones que han llevado a nuevos modelos, como aquellos que incorporan medidas de entropía y lógica difusa. De esta manera, el modelo de media-varianza puede producir soluciones muy variables con incluso pequeños cambios en parámetros como el retorno esperado del portafolio. Como se señala en [11], también puede asignar valores altos a activos de alto riesgo. Esta variabilidad se debe, en parte, a la suposición del modelo sobre la distribución de los retornos esperados de los activos respecto de que, las varianzas y las covarianzas, siguen una distribución normal, y la expectativa de que estos resultados se mantendrán en el futuro.

En este contexto, el nivel de retorno esperado en el modelo refleja las expectativas de los inversores. Es una práctica común realizar análisis de sensibilidad sobre este parámetro, ya que los resultados han demostrado que pequeños cambios en el retorno esperado pueden dar lugar a portafolios significativamente diferentes [12, 13]. Para abordar esto, se propone agregar flexibilidad tratando el retorno esperado como un rango en lugar de un valor fijo.

Esto se logra utilizando funciones de pertenencia como se indica en el capítulo 3, sección 3.2.

El objetivo del modelo de media-varianza es lograr el mayor nivel de retorno posible minimizando la varianza. Sin embargo, el portafolio resultante puede contener activos de alto rendimiento para cumplir con el retorno deseado, lo que también se conoce como una solución escasa [11, 14]. Para esto, es posible contar con herramientas como la diversificación y la maximización de la entropía del portafolio [15], los cuales se pueden utilizar para abordar este problema.

Es común asumir que los retornos siguen una distribución normal al estimar medias, varianzas y covarianzas en finanzas. Sin embargo, los retornos de los activos a menudo no se distribuyen normalmente, presentando sesgo y curtosis, lo que lleva a generar portafolios que parecen eficientes basados en suposiciones normales, pero que son subóptimos cuando se considera su distribución real [16].Además, asumir distribuciones normales también puede resultar en evaluaciones de riesgo demasiado optimistas, permitiendo a los inversores asumir más riesgo del que se dan cuenta, ya que este tipo de análisis no captura adecuadamente eventos inesperados como grandes caídas del mercado [17]. Además, las suposiciones normales pueden afectar negativamente la diversificación del portafolio, ya que las correlaciones estimadas a partir de distribuciones normales pueden diferir de las correlaciones reales en los datos. En base a lo anterior, la hipótesis central para abordar los modelos multiobjetivo propuestos consiste en que el uso de entropía difusa en retornos esperados y varianzas podría ayudar a mejorar el desempeño de los portafolios tanto en rentabilidad como en eficiencia, gracias a u capacidad de poder manejar de mejor manera la incertidumbre y eventos poco probables.

Por otra parte, existen enfoques que utilizan medidas de entropía para abordar algunos de los problemas del modelo de media-varianza; el uso más común es producir portafolios diversificados [11]. En este contexto, [15] entendió la necesidad de mayor diversificación de las soluciones del modelo de media-varianza como soluciones de esquina y analizó cómo diferentes medidas de entropía podrían ayudar con este problema.

Como se indicó en [18], los enfoques difusos manejan mejor las incertidumbres del mundo real en la construcción de portafolios y las medidas de entropía ayudan a la diversificación al evitar soluciones de esquina [15]. La conclusión directa que deriva de estos estudios es considerar múltiples objetivos entre los cuales se puede mencionar al menos medidas de entropía y retorno con el uso de parámetros difusos [18, 19, 20].

En este contexto, una parte primordial de la investigación es probar si la introducción de entropía difusa y lógica difusa produce mejores portafolios que los obtenidos con enfoques conocidos, como el modelo base de media-varianza y los modelos con medidas de

entropía. Para esto, se representa la ganancia de información como una de las funciones objetivo y se controla el parámetro difuso en las restricciones del problema de optimización.

Además, se compara el desempeño de los portafolios a través de la asignación de activos utilizando siete modelos de optimización en diferentes momentos. Posteriormente, se utilizan diversas ventanas de tiempo para calcular el portafolio resultante en cada modelo, y dentro de estos, se utilizan múltiples configuraciones para analizar la relevancia de las funciones objetivo. Luego, se calcula el desempeño de los portafolios para datos fuera de la ventana de tiempo en la que se estima el portafolio. Con todos estos resultados, se realiza un ranking de portafolios mediante un enfoque multicriterio TOPSIS (Technique for order Preference by Similarity to Ideal Solution).

En la segunda etapa, se estudia un enfoque basado en múltiples agentes que operan en un entorno compartido y son entrenados ya sea individualmente o en un modelo multiagente lograndose obtener soluciones más consistentes frente a la incertidumbre del mercado, lo que resulta en una mejor gestión de riesgos y mayores rendimientos.

Para ello, la construcción de portafolios, mediante Aprendizaje reforzado multiagente, se aborda como una extensión al control estocástico a partir de la posibilidad de minimizar simultáneamente la varianza del portafolio mientras se maximiza el retorno esperado directamente en la función de recompensa del agente, permitiendo una gestión de portafolios que es tanto rentable como estable.

Se propone un modelo multiagente donde los agentes aprenden a maximizar sus utilidades, no solo basándose en la información del mercado, sino también en las estrategias de otros agentes. Se considera un ambiente que es competitivo y cooperativo acorde a los mercados financieros, se busca que los agentes desarrollen estrategias de inversión robustas frente a las acciones de otros participantes del mercado intentando diversificar el portafolio, mediante la minimización de la correlación entre los portafolios de los agentes y maximización del retorno junto con él índice de Sharpe del mismo, considerando además penalizaciones por los de movimientos en los portafolios o costos de transacción. Este enfoque representa un avance significativo en la capacidad de los modelos financieros para captar la complejidad de los mercados reales, donde las interacciones entre diferentes actores son cruciales para el comportamiento del mercado.

Una de las contribuciones más importantes de este escenario es la capacidad de modelar situaciones en las que los agentes alcanzan un equilibrio de Nash, donde ninguno tiene incentivos para desviarse unilateralmente de su estrategia,pero deben velar por el objetivo general. De tal manera, que los agentes buscan estrategias de inversión que sean sostenibles frente a las decisiones de otros actores del mercado. Para esto, se considera la coordinación de múltiples agentes que interactúan en un entorno compartido, garantizando la estabilidad

y convergencia de las políticas aprendidas a través de un agente que busca maximizar una función de recompensa global sin perder de vista objetivos individuales. Es un reto considerable, ya que los agentes deben alcanzar un equilibrio que no solo sea localmente óptimo, sino también a nivel global y robusto frente a perturbaciones externas o cambios en las estrategias de otros agentes.

En los mercados financieros, los eventos inesperados y los cambios abruptos en el entorno pueden hacer que las estrategias que funcionan bien en condiciones normales sean ineficaces, o incluso perjudiciales durante periodos de crisis. Por lo tanto, el planteamiento del modelo MARL nace de la necesidad de desarrollar un modelo que no solo optimice el rendimiento bajo condiciones típicas del mercado, sino que también sea capaz de adaptarse rápidamente a nuevas realidades.

Desde el análisis de modelos multicriterio, se da solución a la toma de decisiones, escogiendo el portafolio óptimo de mercado considerando los modelos clásicos como Markowitz pero con variantes dadas por lógica difusa y entropía difusa en modelos multi-objetivos.

En este escenario,se otorga una solución a la selección de portafolios eficientes, mediante una solución que emula un sistema de inversiones mediante un modelo de aprendizaje reforzado multiagente (MARL)y por otra parte, mediante un modelo multiobjetivo con entropía difusa.

De acuerdo a los dos planteamientos, se explora modelos RL uniagentes en conjunto con un modelo entrópico difuso, lo que en esta tesis recibe el nombre de modelo híbrido, donde los portafolios obtenidos por cada agente son vistos como activos de un portafolio y los ponderadores como probabilidades que poseen un grado de pertenencia a un conjunto, optimizando un portafolio de portafolios, mediante las estrategias de un modelo multiobjetivo con entropía difusa.

De esta manera, ambas estrategias Modelo multiobjetivo y Modelo MARL, junto con el modelo híbrido constituyen diferentes aristas de manejo de portafolios en donde la segunda es altamente prometedora y en conjunto con el modelo híbrido otorgan soluciones muy prometedoras e innovadoras en el ámbito de selección y optimización de portafolios de inversión.

1.2. Objetivos

1.2.1. Objetivo general

El objetivo general es construir un sistema de gestión de capital basado en aprendizaje reforzado profundo multiagente (MARL) para la optimización de portafolios de inversión. Un modelo formado por agentes con diferentes objetivos que operan bajo la supervisión de un agente encargado de lograr una optimización tanto a nivel local como a nivel global.

Un escenario complejo donde gobiernan agentes que aprenden en forma continua, compitiendo y colaborando para que gestionados a través del coordinador o supervisor alcancen mayor diversificación, eficiencia y rentabilidad a nivel global, sin perder de vista la optimización de sus objetivos particulares.

1.2.2. Objetivos específicos

Los objetivos específicos de la investigación son:

1. **Creación de un set de portafolios y elección del óptimo**: Mediante un modelo de optimización multicriterio con técnicas de maximización de entropía y aplicaciones de lógica difusa, bajo restricciones presupuestarias se obtiene un set de portafolios y posteriormente mediante un análisis multicriterio TOPSIS se escoge el óptimo. Descrito en el capítulo 3, sección 2, ítem 3.2.6.

2. **Modelar la arquitectura del sistema financiero**: modelación de múltiples agentes, quienes pueden tomar decisiones diferentes y construir carteras en base a sus propios objetivos, recibiendo la misma información de mercado y sus activos.

3. **Modelar la arquitectura del sistema cooperativo**: Inspirado en las estrategias de diversificación de carteras utilizadas en las grandes empresas de inversión y basándose en el hecho de que los agentes pueden ser vistos como un grupo de "inversores" independientes que compiten y cooperan entre sí. Se crea un sistema cooperativo-diversificado, que considera objetivos y costos de transacción para cada agente y que opera en función del cumplimiento de los objetivos individuales sujetos a un interés global modelado por la arquitectura de supervisión.

4. **Modelar la arquitectura de supervisión del sistema multiagente**: Mediante una capa adicional de supervisión que evalúa el rendimiento conjunto de los agentes y ajusta las ponderaciones del portafolio de acuerdo con los objetivos generales del inversor. Esta capa actúa como una "meta-red", coordinando a los agentes de renta fija y variable, a nivel nacional como internacional, para garantizar que las decisiones de cada uno estén alineadas con el rendimiento esperado global. En este

sentido, la meta-red utiliza técnicas de aprendizaje por refuerzo profundo para ajustar de manera dinámica la asignación de capital entre los agentes, basándose en el rendimiento histórico y proyectado de cada segmento del mercado. Además, la meta-red no es un agente independiente; no opera como un agente más dentro del sistema ni toma decisiones de inversión en un segmento particular, sino que ajusta y coordina las asignaciones de capital para optimizar el rendimiento global.

La meta-red actúa como una función de supervisión, observando las actividades de cada agente y ajustando las ponderaciones del portafolio, en función de los resultados históricos y las proyecciones futuras. De esta manera, mientras los agentes se centran en optimizar su segmento de mercado, la meta-red asegura que las decisiones de cada uno mantengan el equilibrio del portafolio global. Estos ajustes buscan minimizar la exposición a riesgos específicos y maximizar la eficiencia del portafolio completo, lo que permite que cada agente trabaje en colaboración hacia el objetivo global, en lugar de operar de manera aislada. La función de la meta-red se representa mediante una función de recompensa global que evalúa el rendimiento conjunto de los agentes.

5. **Conceder solución a la toma de decisiones**: Escogiendo el portafolio óptimo de mercado por medio de un modelo de MARL. El objetivo es considerar los diferentes portafolios creados por los agentes en cada una de sus áreas respectivas, construyendo un portafolio de portafolios, donde la obtención de cada uno deriva de la aplicación de un modelo donde un agente determinado trabaja en un entorno multiagente eligiendo el óptimo de mercado.

6. **Conceder solución híbrida a la toma de decisiones**: Mediante un modelo multicriterio que incorpora entropía difusa, tomando los portafolios de cada agente como un activo con una función de pertenencia entrópica. Representando la ganancia de información como una de las funciones objetivo y controlando el parámetro difuso en las restricciones del problema de optimización.

   Es decir construir una cartera de portafolios optimizada mediante aprendizaje reforzado profundo (RL) y mediante optimización difusa con funciones de pertenencia entrópicas determinar el portafolio compuesto por portafolios, generando un modelo híbrido de optimización que pretende competir con el modelo multiagente, Este objetivo es complementario al global e incorpora una novedosa manera de entender los portafolios de los agentes dada la variabilidad e incerteza de los mercados financieros, donde ahora los portafolios y no los ponderadores son vistos como probabilidades que poseen un grado de pertenencia a un conjunto.

   En otras palabras, la optimización global de portafolios de inversion considerando un modelo multiagente, o un escenario con lógica difusa aplicada a portafolios

resultantes de modelos de RL uniagentes, resulta ser una desafiante y prometedora manera de abordar la toma de decisiones de un único portafolio. En la primera se maximiza tanto localmente como globalmente hasta encontrar una convergencia y en la segunda hay maximizaciones locales sujetas a un modelo de maximización entrópico difuso global.

1.3. Hipótesis

1.3.1. Hipótesis General

Si la construcción y gestión de portafolios se formula como un problema de decisión secuencial resuelto mediante aprendizaje por refuerzo profundo multiagente (MARL), coordinado por una meta-red que ajusta dinámicamente las ponderaciones de los portafolios optimizados localmente por cada agente, bajo objetivos globales, entonces el portafolio de portafolios resultante presenta, en comparación con los enfoques de referencia (equiponderado, media–varianza y RL uniagente), menor correlación, menores costos de transacción totales, mayor rentabilidad y mejor desempeño ajustado por riesgo, dada una cartera definida, costos y horizonte de evaluación para todos los modelos.

1.3.2. Hipótesis específicas

(H1)La introducción de lógica difusa y entropía difusa produce portafolios más rentables y eficientes que los obtenidos con enfoques clásicos, como el modelo base de media-varianza de Markowitz, modelos formados a partir de la entropía del portafolio y modelos que consideren lógica difusa clásica.

(H2) si múltiples agentes, con la misma información y objetivos propios, construyen portafolios de manera simultánea, entonces se obtiene mayor diversidad de señales y mejores combinaciones globales que con un agente único, dado universo, reglas y costos equivalentes.

(H3) si la función global penaliza la correlación entre agentes e incorpora costos por agente, entonces disminuye la correlación entre portafolios y mejora la diversificación del portafolio total, dado el mismo esquema de rebalanceo.

(H4) si una meta-red reasigna capital entre agentes conforme a objetivos y desempeño global, entonces aumenta la rentabilidad conjunta, el sharpe global y la robustez frente a cambios de régimen, dado datos, costos y restricciones compartidas.

(H5) si la selección dinámica del portafolio de portafolios se aprende mediante aprendizaje por refuerzo profundo multiagente, entonces se supera a reglas determinísticas de

elección, dado el mismo conjunto de candidatos y restricciones.

(H6) si la decisión final integra multicriterio con entropía difusa, tratando cada portafolio-agente como activo difuso, entonces las decisiones son más estables y mantienen o mejoran el desempeño ajustado por riesgo frente a usar solo aprendizaje por refuerzo o solo multicriterio, dado el mismo universo y costos.

1.4. Estructura del informe

La presente tesis se organiza en un total de seis secciones, siendo la siguiente sección una presentación de la revisión de la literatura, y posicionando el trabajo dentro de la literatura actual de selección de portafolios. Por su parte, en la sección 3 se describen los modelos de optimización y el modelo de AR multiagente. Mientras tanto, la sección 4 describe los datos considerados para los experimentos numéricos y los detalles de la metodología aplicada. Luego, la sección 5 muestra los resultados obtenidos, comentados y discutidos. Finalmente, dentro de la sección 6 se presenta el análisis final y las conclusiones.

Capítulo 2

Revisión bibliográfica

El campo de la optimización de portafolios es amplio y se origina con el trabajo clásico de Markowitz [1]. Desde entonces, el número de artículos en el área ha seguido creciendo, con muchas investigaciones en el campo, como las de [10] mencionadas en modelos deterministas o las de [21] orientadas a métodos de optimización robusta.

A partir de lo anterior, se introduce la Teoría Moderna de Portafolios (MPT), una teoría de inversión basada en la idea de que los inversores adversos al riesgo pueden construir portafolios para optimizar o maximizar los rendimientos esperados, basados en un nivel dado de riesgo de mercado y enfatizando que el riesgo es una parte inherente de mayores recompensas [1]. La MPT asume que los inversores son adversos al riesgo, lo que significa que, dado dos portafolios que ofrecen el mismo rendimiento esperado, preferirán el menos arriesgado. Así, un agente asumirá un mayor riesgo solo si es compensado con rendimientos esperados más altos. Inversamente, un agente que desee mayores rendimientos esperados debe aceptar más riesgo. La compensación exacta será la misma para todos los inversores, pero evaluarán la compensación de manera distinta según sus características individuales de aversión al riesgo. Una estrategia común empleada por un agente es reducir el riesgo del portafolio manteniendo combinaciones de instrumentos que no estén perfectamente correlacionados positivamente. Si todos los pares de activos tienen correlaciones nulas son perfectamente no correlacionados. Además, la varianza del rendimiento del portafolio, medida de riesgo del mismo, es la suma sobre todos los activos del cuadrado de la fracción del portafolio mantenida en el activo multiplicada por la varianza del rendimiento del activo, y la desviación estándar del portafolio es la raíz cuadrada de dicha suma.

La frontera eficiente es una piedra angular de la teoría moderna de portafolios y es la línea que indica la combinación de inversiones que proporcionará el nivel más alto de rendimiento para el nivel más bajo de riesgo. Cuando un portafolio cae a la derecha de la

frontera eficiente, posee un mayor riesgo relativo a su rendimiento predicho. Cuando cae por debajo de la pendiente de la frontera eficiente, ofrece un menor rendimiento relativo al riesgo [1]. Es una de las teorías económicas más importantes e influyentes que trata sobre finanzas e inversiones.

A continuación, se mencionan los cinco supuestos básicos que son los fundamentos sobre los que se construye la MPT [1]:

1. El rendimiento esperado y la varianza son los únicos parámetros que afectan la decisión del inversor.

2. Los inversores son generalmente racionales y adversos al riesgo. Son completamente conscientes de todos los riesgos en la inversión y toman posiciones basadas en la determinación del riesgo, exigiendo un mayor rendimiento por aceptar una mayor volatilidad.

3. No existen costos de transacción para la compra y venta de valores.

4. Todos los inversores tienen las mismas expectativas con respecto al rendimiento esperado, la varianza y la covarianza.

5. El análisis se basa en un modelo de inversión de un solo período.

A pesar de su importancia teórica, los críticos de la MPT cuestionan si es una herramienta de inversión ideal, ya que su modelo de los mercados financieros no coincide con el mundo real en muchos aspectos. Las medidas de riesgo, rendimiento y correlación utilizadas por la MPT se basan en valores esperados, lo que significa que son afirmaciones matemáticas promedio.

Por su parte, el modelo de media-varianza considera a priori la homocedasticidad de la varianza de la serie, un hecho que contradice uno de los hechos estilizados de las series financieras, que es la frecuente heterocedasticidad, es decir, la varianza generalmente sufre cambios sistemáticos durante el tiempo de estudio. En la práctica, los inversores deben sustituir predicciones basadas en mediciones históricas de los retornos de los activos y su volatilidad por estos valores en las ecuaciones. Es por esto dichos valores esperados suelen no tener en cuenta nuevas circunstancias que no existían cuando se generaron los datos históricos, lo que introduce incertidumbre en el problema, modelada a través de la lógica difusa [12, 13].

En este contexto, se han utilizado muchos métodos para optimizar portafolios, sin embargo la atención que han recibido los métodos difusos no ha sido sistemática, la mayoría de las investigaciones consideran métodos paramétricos con diversificación de

carteras, pero el tratamiento de incertidumbre o caídas abruptas de mercado mediante lógica difusa no es amplio al nivel de los estudios mencionados. Una excepción es el estudio de [19],que no ha sido actualizado desde entonces, pero ha dado lugar a muchos otros desarrollos alternativos. A continuación, describiremos otros trabajos dentro de la optimización de portafolios utilizando métodos difusos.

Enfoques como [22] y [23] han aplicado métodos difusos para modelar los rendimientos, considerando su varianza como la variable difusa e identificando la variabilidad de los rendimientos como una forma de incertidumbre. Estos modelos hicieron suposiciones sobre la forma de la varianza o consideraron escenarios de varianza determinista. Posteriormente, los investigadores se centraron en mantener la visión difusa de la incertidumbre de los rendimientos pero modelaron la dinámica temporal del problema. Artículos como [24], [25] y [26] trataron la optimización de portafolios con variación temporal en diferentes circunstancias, los cuales tratan desde la acumulación de riesgos de la varianza, costos de transacción y grados de diversificación [24], considerando directamente la serie temporal de rendimientos como un sistema difuso [27], hasta aplicar la incertidumbre a los rendimientos y la liquidez de los activos dentro de la serie temporal [26], y luego por [28]. En todos estos estudios, el enfoque es el estudio de los rendimientos a lo largo del tiempo como un sistema difuso, a diferencia de este trabajo, en el que se considera un sistema difuso bidimensional sobre la superficie media-varianza, combinando la fuente de incertidumbre.

Otro enfoque, que se basa en la utilización de métodos difusos en modelos multiobjetivos. El trabajo reciente de [29] explora cómo se pueden reconciliar tres objetivos diferentes al optimizar portafolios utilizando lógica difusa. [30] compara el rendimiento de las acciones y la situación financiera general de la empresa para construir un portafolio siguiendo la lógica difusa. [31] modela los rendimientos como un sistema difuso, pero incluye objetivos difusos en el proceso de optimización, incluyendo restricciones de credibilidad perceptible en la construcción del portafolio, un enfoque similar al trabajo de [32], quienes consideran la credibilidad dentro del espacio de retorno-liquidez. Siguiendo este enfoque de credibilidad, el trabajo de [33] considera rendimientos, volatilidad, sesgo y curtosis dentro de un marco difuso, pero se enfoca en cada uno como un objetivo en lugar de una fuente de incertidumbre, como lo hacemos en nuestro trabajo.

Además, otros enfoques en el área se han centrado, por ejemplo, en reducir la variabilidad en la decisión, como el enfoque de semi-entropía de [27], que se centra en la incertidumbre a la baja. El impacto de la racionalidad en las decisiones utilizando entropía difusa fue explorado recientemente por [34], quienes modelaron restricciones en los portafolios a partir de las expectativas de racionalidad en los agentes. Este trabajo sigue la línea más tradicional de maximizar el rendimiento ajustado al riesgo del portafolio al incluir explícitamente la incertidumbre de la superficie media-varianza, como se explica

en el siguiente capítulo.

Por otro lado, en el contexto de un sistema de apoyo a la decisión para la clasificación de inversiones en acciones de empresas basadas en la inversión en valor, el enfoque TOPSIS (Técnica para el Orden de Preferencia por Similitud con la Solución Ideal) puede ser una herramienta valiosa.

En este contexto, el método utilizado por este estudio como sistema de apoyo a la decisión de clasificación de acciones es el procedimiento TOPSIS, entre muchos otros existentes, por su sencillez y menos subjetividad que otros métodos como AHP o ELECTRE, descrito en el capítulo 3, sección 4.2.3. El principio geométrico del algoritmo TOPSIS establece que la alternativa seleccionada debe estar ubicada a la máxima distancia de la solución ideal negativa y lo más cerca posible de la solución ideal positiva. Para determinar la proximidad relativa de una alternativa a la solución óptima, se utiliza la distancia euclidiana como un método de toma de decisiones multicriterio, el cual clasifica las posibles soluciones de acuerdo con su similitud con la solución óptima. Varios estudios han destacado la aplicabilidad de TOPSIS en varios dominios.

En el campo del aprendizaje no supervisado los primeros estudios abordaron el análisis de componentes principales (PCA) [35], que descompone los datos multidimensionales en un conjunto de variables linealmente no correlacionadas. De esta manera, la primera de esas variables o componente principal explica la mayor variación en los datos, y todas las siguientes variables se ordenan teniendo la varianza máxima mientras son ortogonales a la variable anterior. El primer componente principal sirve como una aproximación del mercado, por lo tanto, elegir el segundo y otros componentes no correlacionados con las estrategias de mercado es lo que quieren la mayoría de los inversores.

Otro modelo sin supervisión utilizado es el de paridad de riesgo [36]. También se plantea como un problema de optimización, en el que asignamos más el riesgo que los recursos de capital. El problema de este enfoque es que madura cuando las carteras son de tamaño muy grande: si representamos las conexiones entre activos geométricamente es una sobrecomplicación en el mundo de las jerarquías. La solución está nuevamente en el aprendizaje no supervisado, pero con la explotación de algoritmos de agrupamiento jerárquico aplicado a la matriz de covarianza [37]. Después de encontrar grupos de activos, se puede reasignar el riesgo sobre ellos de forma recursiva.

Luego, considerando inteligencia artificial aplicada en el área, aparecen modelos de redes neuronales usados para la optimización de carteras. Las redes neuronales convolucionales y las redes neuronales recurrentes son arquitecturas de redes neuronales profundas utilizadas en una variedad de dominios [38] y que demuestran un desempeño exitoso en el procesamiento de datos multidimensionales (imágenes y vídeos) [38]. Como arquitectura

profunda discriminativa reducen la complejidad del aprendizaje mediante convolución, siendo utilizadas para el reconocimiento y tareas de clasificación en multimedia, además de ser eficientes en los análisis de datos financieros.

Tsantekidis y col. [39] propusieron un esquema de redes neuronales convolucionales (CNN) profundas para la predicción de movimientos de precios utilizando un gran conjunto de datos de transacciones de alta frecuencia. Del mismo modo, Chen [40] formularon un modelo de CNN profundo que utiliza la representación de características planas para el análisis y la predicción de los precios de las acciones. Su esquema logró una precisión del 57,88 % en la predicción de tres categorías.

Las Redes Neuronales Recurrentes (RNR) [41] difieren de las redes pre-alimentadas, ya que poseen un circuito de retroalimentación de las decisiones pasadas, guardando la información en su memoria. Son eficaces en el pronóstico de fenómenos naturales [42] y [43], así como en el procesamiento del lenguaje y finanzas [44]. Di Persio utilizó diversos tipos de redes neuronales para la predicción del precio de acciones. Ghorbani [45] pronosticó precios de cierre de acciones previa reducción de la dimensión por medio de análisis de componentes principales (PCA). Wang [46],obtiene un principio de máximo para problemas de control óptimo recursivo estocástico parcialmente observados bajo el supuesto de que los dominios de control no son necesariamente convexos y desarrolló una arquitectura de Elman para la predicción de índices de precios en los mercados de valores. Rout [47]propone un nuevo modelo de pronóstico no lineal que integra una red neuronal de enlace funcional con una red neuronal de función de base radial para mejorar el rendimiento de la predicción, utiliza una RNR de baja complejidad con aprendizaje en red y evolutivo para pronosticar el índice S&P500 e investiga el rendimiento comercial y estadístico de todos los modelos en una simulación de pronóstico de los tipos de cambio entre el dólar estadounidense y otras cuatro monedas principales, euro, rupia india, dólar canadiense, dólar australiano, etc. Su esquema demostró una varianza muy baja y buen desempeño en la predicción de la volatilidad, superando a las redes neuronales prealimentadas en predicción de series de tiempo financieras.

Por un lado, el estudio de [48] presenta un sistema de trading que permite orientar a un agente en la decisión de cuándo entrar o salir del mercado a través de indicadores difusos, además de cuantificar los activos en las operaciones, utilizando soft computing, específicamente algoritmos genéticos para la optimización de los parámetros del sistema. El efecto de utilizar un sistema difuso es suavizar las pérdidas, además de una mayor probabilidad de obtener beneficios y de que estos sean más altos.

Por otro lado, el estudio de [49] utiliza el índice bursátil alemán DAX-30 en la aplicación de una red neuronal difusa híbrida para predecir la dirección de compra o venta del índice.

Se implementan diferentes modelos y una combinación lineal de ellos obtenida a través de un análisis de factores. Los resultados muestran que la reducción de la dimensión a través del análisis factorial genera estrategias más rentables y menos arriesgadas.

Además, el estudio de [50] presenta el estudio de una acción volátil de la Bolsa de Valores de Colombia la cual es analizada usando un modelo estadístico ARIMAX (Modelo autorregresivo integrado de media móvil con entrada exógena) y un SONFS (sistema neuro difuso auto organizado). Estos métodos son comparados teniendo en cuenta tres características: el menor EMA (Error Medio Absoluto),el residual entendido como un proceso de ruido blanco (evaluado mediante seis pruebas) y el AIC (criterio de información de Akaike); así se elige el model oque mejor se ajusta para la predicción de la serie.

Continuando los avances en los últimos años, la gestión de portafolios ha experimentado una notable transformación gracias a la incorporación de técnicas avanzadas de aprendizaje automático, particularmente el aprendizaje reforzado (RL). Este enfoque ha demostrado ser efectivo para manejar la complejidad inherente a los mercados financieros, donde la toma de decisiones óptima es crucial en entornos dinámicos y estocásticos. Los primeros trabajos en este campo, como el de Moody y Saffell [51], fueron pioneros en mostrar cómo los agentes de RL podían aprender estrategias de trading sin necesidad de modelos predictivos explícitos. Su trabajo introdujo el concepto de aprendizaje directo mediante refuerzo, donde el agente interactuaba directamente con el mercado para maximizar el retorno sobre la inversión, marcando un hito en la aplicación del RL en finanzas.

Con la evolución de las redes neuronales profundas, se dio un paso significativo hacia adelante en el uso del RL en la gestión de portafolios. Mnih [52] logró integrar estas redes con algoritmos de RL, creando lo que hoy se conoce como Deep Reinforcement Learning (DRL). Este avance permitió que los agentes financieros pudieran manejar grandes volúmenes de datos de mercado, capturando patrones complejos y mejorando la capacidad de generalización de los modelos. La capacidad de DRL para aprender políticas de decisión en entornos de alta dimensionalidad revolucionó la forma en que se abordan los problemas de trading algorítmico, superando las metodologías tradicionales basadas en reglas estáticas y modelos lineales.

Otro enfoque relevante dentro del RL en finanzas es la extensión al control estocástico, donde Moody y Saffell [51] exploraron la posibilidad de minimizar simultáneamente la varianza del portafolio mientras maximizaban el retorno esperado. Este enfoque introdujo la idea de incluir el control del riesgo directamente en la función de recompensa del agente, permitiendo una gestión de portafolios que es tanto rentable como estable. Esta línea de investigación ha sido fundamental para el desarrollo de estrategias de inversión que integran la gestión del riesgo como un componente central, diferenciándose de los enfoques

que solo buscan maximizar el rendimiento sin considerar la estabilidad.

Además, se incluyen modelos básicos que resuelven el problema con un agente, una función de recompensa de retorno y un algoritmo de minimización de pérdidas, hasta modelos profundos con la utilización de redes neuronales como agentes, funciones de recompensa orientadas a la eficiencia en la obtención de portafolios óptimos e introducción de técnicas multi-agentes. Los estudios de [53, 54, 55, 56, 57, 58] utilizan aprendizaje reforzado para predicción de precios hasta gestión de portafolios.

En el estudio de [59] se aplica aprendizaje reforzado en conjunto con aprendizaje difuso, y en el estudio de [60] se utiliza aprendizaje reforzado recurrente logrando alta eficiencia en la gestión de activos. En [61], se forma una cartera de criptomonedas mediante aprendizaje reforzado profundo, donde los experimentos realizados demostraron que el modelo diseñado basado en RL logra retornos del orden de 10 veces el valor en períodos de 1,8 meses. Además, el estudio de [62] propone un sensor de navegación basado en lógica difusa y aprendizaje reforzado; las lecturas del sonar son codificadas en distancias en conjuntos difusos, incorporando modificaciones al algoritmo RL para integrar lógica difusa. Del mismo modo, los estudios de [63, 64, 65, 66, 67] tratan distintos tipos de estrategias para comprensión del mercado, destacando las más recientes el entrenamiento mediante aprendizaje reforzado aplicados a portafolios o a predicción de precios, incluyendo entrenamiento con doble Q-learning y otras técnicas de aceleración de tiempos de entrenamiento.

Un avance clave en la aplicación del RL a la gestión de portafolios fue presentado por Li [68], quienes utilizaron DRL para optimizar dinámicamente la composición de un portafolio en función de las condiciones del mercado. Su investigación demostró que los modelos de RL podían aprender a equilibrar eficazmente el riesgo y el retorno, adaptándose rápidamente a los cambios en el entorno financiero. Este enfoque no solo se centró en maximizar el rendimiento, sino que también incorporó medidas de riesgo, optimizando la robustez del portafolio frente a la volatilidad del mercado. Este desarrollo representó un avance significativo sobre los modelos estáticos de optimización de portafolios, que no pueden adaptarse rápidamente a las condiciones cambiantes del mercado.

El estudio de [69] utiliza aprendizaje reforzado (AR) con un modelo de redes neuronales convolucionales (CNN) y recurrentes (RNR) como posibles agentes.Este artículo presenta un novedoso enfoque de dos etapas que integra el aprendizaje profundo con la optimización de carteras. En la primera etapa, desarrolla un modelo de predicción de tendencias bursátiles para la preselección de acciones, denominado modelo AGC-CNN, que aprovecha una red neuronal convolucional (CNN), un mecanismo de autoatención, una red convolucional de grafos (GCN) y vecinos más cercanos k-recíprocos (NN k-recíprocos),

Utiliza una matriz de precios históricos en una ventana de tiempo diaria considerada como serie de tiempo que caracteriza al ambiente en t como el Estado(t), utilizando el retorno como función de recompensa. La evaluación del desempeño del modelo muestra un portafolio con un desempeño más alto en términos de rentabilidad, incrementando el retorno en un 39 % y aminorando las caídas en un 13,7 % en comparación a un modelo AR con el retorno como función única de recompensa.

Adicionalmente, en el estudio de [70] se desarrolla un sistema de gestión de portafolio para posiciones largas y cortas de inversión. De esta manera, se integran ambas y se consideran las características de un EMN (Equity Market Neutral), esto con el fin de dar protección al portafolio frente al riesgo.

Entre las diferentes estrategias de portafolio, Equity Market Neutral (EMN) [71, 72], es una estrategia de cobertura que disminuye la exposición al riesgo al equilibrar la tenencia de acciones relativamente fuertes con la venta de acciones relativamente débiles [73]. Todo el capital procedente de posiciones cortas se reinvierte en el mercado y cubre todas las posiciones largas, mientras que las posiciones largas y cortas equivalentes se utilizan para reducir el riesgo sistémico en el mercado. Teóricamente, entonces, no impone ningún capital de inversión, y el dinero de inversión neto es cero.

En el estudio de [74] se aplica Deep Q-Network con un aproximador de funciones de red neuronal convolucional, que toma imágenes de gráficos bursátiles como entrada para realizar predicciones del mercado bursátil global. El modelo no solo produce ganancias en el mercado bursátil del país cuyos datos se usaron para entrenar, sino que también produce ganancias en los mercados bursátiles globales,las carteras construidas con base en el resultado de del modelo generalmente producen un rendimiento de alrededor del 0,1 al 1,0 por ciento por transacción antes de los costos de transacción en los mercados bursátiles de 31 países

En el estudio de [75] se implementa un sistema de Gestión de Portafolio basado en aprendizaje reforzado con múltiples agentes (MAPS). Uno de los primeros trabajos del tipo aprendizaje reforzado ensamblado [76], consistente en el apilamiento de múltiples agentes de Deep Q-Network, que otorgan soluciones confiables y explícitas implementando la política del gradiente descendente [61]. De esta manera, MAPS es un sistema en el que cada agente crea su propia cartera.

El aprendizaje reforzado multiagente (MARL) ha surgido como una extensión natural del AR, proporcionando un marco para modelar la interacción entre múltiples actores en un entornofinanciero, [77]. Yang y Wang [78] exploraron la aplicación del MARL en la optimización de inventarios en cadenas de suministro, sentando un precedente para su uso en finanzas. Aunque su trabajo no estaba directamente enfocado en la gestión de

portafolios, estableció las bases para comprender cómo los agentes pueden interactuar en un entorno competitivo. Este marco es particularmente útil en los mercados financieros, donde las decisiones de un agente pueden afectar significativamente las oportunidades y riesgos enfrentados por otros.

En el ámbito específico de la gestión de portafolios, Lee y Kim [75] adaptaron el MARL para modelar la interacción entre múltiples inversores o gestores de fondos. En su modelo, los agentes aprendían a maximizar sus utilidades no solo basándose en la información del mercado, sino también en las estrategias de otros agentes. Este trabajo subrayó la importancia de la competencia en los mercados financieros, mostrando cómo los agentes pueden desarrollar estrategias de inversión robustas frente a las acciones de otros participantes del mercado. Este enfoque representa un avance significativo en la capacidad de los modelos financieros para captar la complejidad de los mercados reales, donde las interacciones entre diferentes actores son cruciales para el comportamiento del mercado. En [79] se propone un marco que combina múltiples agentes con diferentes preferencias de inversión para integrar estrategias mediante una estructura jerárquica y mejorar el rendimiento de las estrategias de trading. Los resultados experimentales con conjuntos de datos reales muestran que las estrategias de trading desarrolladas con este marco han demostrado superioridad sobre los enfoques de un solo agente y otros algoritmos de referencia.

Una de las contribuciones más importantes del MARL es la capacidad de modelar situaciones en las que los agentes alcanzan un equilibrio. Esto es especialmente relevante en finanzas, donde los agentes buscan estrategias de inversión que sean robustas frente a las decisiones de otros actores del mercado. Aunque la investigación en este campo está en sus primeras etapas, el potencial para comprender y modelar mejor las dinámicas del mercado es inmenso. A medida que se desarrollan nuevas técnicas y algoritmos, es probable que aplicaciones más sofisticadas del MARL en la optimización de portafolios y la gestión de riesgos, ofrezcan soluciones que sean más eficientes y estables en entornos financieros volátiles.

Uno de los principales desafíos pendientes en el MARL es la alta complejidad computacional asociada con la coordinación de múltiples agentes que interactúan en un entorno compartido. Esta complejidad no solo dificulta el desarrollo de políticas óptimas en tiempo real, sino que también plantea problemas de escalabilidad cuando se intenta aplicar estos modelos a mercados financieros de gran tamaño y con múltiples activos. Además, garantizar la estabilidad y convergencia de las políticas aprendidas es un reto considerable, ya que los agentes deben alcanzar un equilibrio que no solo sea localmente óptimo, sino también robusto frente a perturbaciones externas y cambios en las estrategias de otros agentes.

Otro desafío crítico es la robustez de las estrategias aprendidas en el MARL. En los mercados financieros, los eventos inesperados y los cambios abruptos en el entorno pueden hacer que las estrategias que funcionan bien en condiciones normales sean ineficaces o incluso perjudiciales durante periodos de crisis. Por lo tanto, desarrollar modelos que no solo optimicen el rendimiento bajo condiciones típicas del mercado, sino que también sean capaces de adaptarse rápidamente a nuevas realidades, es un área clave de investigación en la que se están centrando los estudios más recientes y esta tesis.

En conclusión, la investigación sobre aprendizaje reforzado en la gestión de portafolios ha avanzado desde sus primeras aplicaciones, pasando del RL aplicado a estrategias de trading individuales hasta el uso de complejos sistemas multiagente que modelan la interacción de actores en el mercado. Los desarrollos en Deep Reinforcement Learning (DRL) han permitido a los agentes manejar la complejidad de los datos financieros de manera más efectiva, mientras que la adopción de MARL ha abierto nuevas posibilidades para modelar la competencia y la cooperación en los mercados. Sin embargo, a medida que la investigación continúa, queda claro que hay importantes desafíos por superar, especialmente en términos de complejidad computacional, estabilidad de las políticas, y robustez frente a la volatilidad del mercado. A pesar de estos retos, es indudable que el AR y el MARL seguirán desempeñando un papel crucial en la innovación de las estrategias de gestión de portafolios, ofreciendo herramientas cada vez más sofisticadas para optimizar la toma de decisiones en entornos financieros dinámicos y complejos.

Capítulo 3

Fundamentos teórico conceptuales en el modelamiento

Este capítulo trata de una descripción de los modelos formulados en esta investigación, a partir de la descripción del problema de selección de portafolio en términos generales y el manejo de las AFP (Administradoras de fondos de pensiones) respecto al manejo actual, para continuar con los contextos teóricos de modelación correspondientes.

Primero, se consideran los modelos clásicos de optimización paramétrica, contexto teórico general que sustenta la metodología abarcada en esta tesis para la construcción de nuevos modelos y sus portafolios desde la optimización multicriterio.

Segundo, se construye un marco general para la comprensión de los modelos de aprendizaje reforzado profundo multiagente, el cual sustenta la metodología utilizada para la construcción de portafolios desde el punto de vista de las técnicas de inteligencia artificial inmerso en un sistema de inversiones, particularmente las AFP chilenas.

Finalizando con la formulación de la utilización de ambas metodologías en un escenario complejo de selección de portafolios.

3.1. El problema

3.1.1. Planteamiento

En términos matemáticos, la optimización de portafolios corresponde a un problema donde las variables de decisión son variables continuas. Lo que se debe decidir es, dado un conjunto fijo de m productos financieros (en este caso, índices y/o acciones), la proporción de dinero que se debe destinar a cada uno en cada instante de tiempo para maximizar la métrica de desempeño elegida.

Cabe señalar que, además de las m acciones escogidas, se debe considerar un "cash" o monto disponible para invertir, es decir, la moneda base, que se elige en virtud de su estabilidad sobre el resto. Así, todos los índices o acciones están expresados en unidades de la moneda base, y por ende, esta última tiene un precio constante e igual a 1 desde el punto de vista de los modelos. De esta manera, la moneda base sirve como opción segura si el tomador de decisiones considera que no es adecuado invertir en alguno de los índices o acciones disponibles.

Una vez escogida una cantidad m de índices o acciones, para un instante de tiempo t, se debe obtener el vector $Y_t \in \mathbb{R}^{m+1}$ a partir de cierta información de entrada X_t. El vector Y_t indica la distribución del capital en los productos escogidos para ese instante de tiempo y debe cumplir con las siguientes restricciones para que la inversión sea factible:

$$0 \leq y_{i,t} \leq 1 \quad \forall\, y_{i,t} \in y_t \tag{3.1}$$

$$\sum_{i=1}^{m+1} y_{i,t} = 1 \tag{3.2}$$

$$r_t = \frac{p_t}{p_{t-1}} = \left(1, \frac{p_{1,t}}{p_{1,t-1}}, \ldots, \frac{p_{m,t}}{p_{m,t-1}}\right)^T \tag{3.3}$$

El valor del portafolio se define de manera aleatoria. Entonces, se tiene:

$$\tilde{w_t} = \frac{\mathbf{r}_t \odot w_{t-1}}{\mathbf{1}^T(\mathbf{r}_t \odot w_{t-1})} \tag{3.4}$$

Donde el producto punto en el numerador indica la multiplicación elemento a elemento, siendo el vector de "valores finales"de cada activo tras aplicar los retornos a los pesos anteriores (producto elemento a elemento) y el denominador, compuesto por un "1", correspondiente a un vector columna de unos multiplicado por todos los componentes

del vector siguiente, de tal manera que representa la suma total del valor de la cartera tras el rebalanceo, es decir, el valor total de la cartera después de los retornos. De esta manera, el factor residual se puede ajustar en cada instante de tiempo según se muestra en la siguiente ecuación:

$$\mu = c \sum_{i=1}^{m} |w'_{i,t} - w_{i,t}| \tag{3.5}$$

Donde:

- μ: Factor residual o factor de ajuste asociado a los cambios en la composición del portafolio.
- c: Constante de proporcionalidad o escala, que puede representar, por ejemplo, el costo unitario o penalización por cambiar la proporción de activos.
- m: Número total de activos considerados en el portafolio.
- $w_{i,t}$: Peso o proporción del activo i en el portafolio en el tiempo t antes del ajuste.
- $w'_{i,t}$: Nuevo peso o proporción del activo i en el portafolio en el tiempo t después del ajuste o reequilibrio.
- $|w'_{i,t} - w_{i,t}|$: Cambio absoluto en la proporción del activo i entre el estado previo y el ajustado.

A partir de aquí, se puede determinar el valor final del portafolio alcanzado al final del tiempo $t_f + 1$. El capital destinado a cada acción no puede ser negativo y tiene que ser menor o igual que el total del mismo, además de que la suma de todo el capital distribuido en las diferentes acciones debe ser igual al total del mismo. De esta manera, el valor final del portafolio es:

$$P_f = P_0 \exp\left[\sum_{t=1}^{t_f+1} R_t\right] = P_0 \prod_{t=1}^{t_f+1} (r_t \cdot w_{t-1} \cdot \mu_t) \tag{3.6}$$

Donde :

- P_f: Valor final del portafolio al tiempo t_f.
- P_0: Valor inicial del portafolio.
- R_t: Rendimiento logarítmico (log-return) del portafolio en el periodo t.

- r_t: Retorno simple del portafolio en el periodo t.
- w_{t-1}: Vector de pesos o proporciones asignadas a los activos en el portafolio en el periodo $t-1$.
- μ_t: Factor de ajuste o residual en el periodo t, que puede incluir costos de transacción, penalizaciones o modificaciones de rendimiento.
- t_f: Último periodo de la serie temporal considerada.

3.1.2. Supuestos del Problema

En este contexto cabe mencionar que se realizan dos supuestos generales con respecto al mercado:

- **S1 — Nulo impacto en el mercado.** Dado el tamaño de las posiciones y el volumen negociado, se asume que los órdenes de magnitud del modelo no alteran el precio, es decir no hubiera tenido influencia en el mercado.
- **S2 — Nulo deslizamiento.** Dada la liquidez, se asume ejecución al precio cotizado.Dicho de otro modo, el mercado cuenta con una liquidez suficiente para que cada operación se ejecute de manera instantánea al precio en el que fue ordenada.

3.1.3. Manejo y gestión de portafolios en AFPs chilenas

El mercado chileno administrado por las Administradoras de Fondos de Pensiones (AFP) distribuye sus inversiones principalmente en cuatro áreas clave: renta fija nacional, renta fija internacional, renta variable nacional y renta variable internacional. Estas áreas están interrelacionadas y juegan un rol complementario en la construcción de un portafolio diversificado. A continuación, se describen cinco puntos que explican esta relación.

En términos de riesgo y rendimiento del portafolio, es posible notar que cada área contribuye de manera diferente al perfil:

1. **Renta Fija Nacional**: Ofrece estabilidad y bajos niveles de riesgo, pero con rendimientos moderados en el mercado chileno. En este perfil los instrumentos comunes de renta fija en Chile:
   a) **Bonos del Banco Central de Chile**: Emisiones del banco central con diferentes plazos y tasas de interés.
   b) **Letras hipotecarias**: Instrumentos emitidos por bancos para financiar hipotecas.

   c) **Bonos corporativos**: Deuda emitida por grandes empresas chilenas.

   d) **Bonos del tesoro**: Deuda soberana emitida por el gobierno chileno.

2. **Renta Fija Internacional**: Aporta diversificación geográfica y exposición a monedas extranjeras, con rendimientos mayores a la renta fija nacional. El índice Global Aggregate es un referente clave en estos mercados, diseñado para medir el desempeño de los bonos globales con su grado de inversión. Este índice proporciona una visión integral del mercado de bonos globales, incluyendo emisiones en dolares (USD),Euros (EUR), Yen japonés (JPY), Libra estrlina (GBP), entre otras divisas. La composición del índice equilibra bonos de distintos plazos, asegurando una duración intermedia y una diversificación óptima para minimizar riesgos específicos. Su uso es fundamental en la gestión de carteras de renta fija, y como benchmark(referente) en fondos globales, permitiendo a inversionistas institucionales evaluar el desempeño de sus estrategias frente a un mercado representativo de renta fija a nivel mundial. Además, otros instrumentos son:

   a) **Bonos soberanos internacionales**: Emitidos por gobiernos extranjeros.

   b) **Bonos corporativos internacionales**: Emitidos por grandes corporaciones globales.

3. **Renta Variable Nacional**: Introduce mayor potencial de rendimiento, pero con más volatilidad, compuesto por diversos tipos de activos nacionales, como:

   a) **Acciones**: Representan la propiedad parcial de una empresa. Los inversionistas pueden recibir dividendos y obtener ganancias (o pérdidas) de capital al vender sus acciones.

   b) **Fondos de inversión**: Fondos mutuos que invierten en un conjunto de acciones nacionales.

4. **Renta Variable Internacional**: Ofrece acceso a mercados globales y mayores oportunidades de crecimiento, pero con mayor exposición a riesgos globales. La cartera proviene de variadas acciones que presentan: diversificación geográfica, mayor exposición a través de acceso a mercados desarrollados y emergentes y exposición a fluctuaciones del tipo de cambio en condiciones macroeconómicas internacionales.

El balance entre estas áreas permite optimizar el rendimiento ajustado por riesgo. Del mismo modo, la diversificación entre estas áreas ayuda a reducir el riesgo general del portafolio.

- La baja correlación entre renta fija y renta variable permite suavizar las pérdidas cuando los mercados de acciones caen, dado que los bonos tienden a ser más estables.
- La renta fija y variable nacional e internacional poseen correlación ayudando a mitigar los riesgos de exposición concentrada en un solo mercado o sector. Sin embargo, en términos de hedge contra la inflación y el riesgo cambiario, ciertas áreas del portafolio actúan como cobertura contra la inflación y las fluctuaciones cambiarias:
- A nivel individual los bonos en renta fija nacional, indexados a la UF, protegen contra la inflación local.

Además, en términos de exposición a ciclos económicos, es posible notar que las áreas responden de manera diferente a estos:

- En momentos de crecimiento global, la renta variable (nacional e internacional) puede experimentar grandes rendimientos.
- En períodos de incertidumbre o recesión, la renta fija (nacional e internacional) tiende a ser el refugio seguro para los inversores.

Esta adaptabilidad permite ajustar la exposición del portafolio a los ciclos económicos de Chile y del mundo. Se pronuncia una complementariedad en la estrategia de gestión activa, considerando que en tiempos de volatilidad interna, las AFPs pueden aumentar su exposición a renta fija y variable internacional para mitigar riesgos locales.En cambio, durante períodos de crecimiento local, los gestores pueden enfocarse en incrementar su exposición a renta variable nacional para capturar el crecimiento económico chileno.

Esta flexibilidad permite a los gestores ajustar la composición del portafolio según las condiciones de mercado locales y globales.

3.2. Modelos de optimización paramétrica

Los modelos de optimización paramétrica abarcan desde el modelo clásico de Markowitz (modelo 1), fundamento teórico para la mayoría de los estudios en ámbitos de selección de portafolios y se realizan 6 variantes con medidas entrópicas: modelo 2 que maximiza la entropía de Shannon con rendimiento y varianza como restricciones, mientras que el modelo 3 minimiza la varianza y maximiza la entropía en una única función objetivo. Mientras tanto, el modelo 4 es un modelo multicriterio con maximización de entropía de Shannon y parámetros difusos en la media y varianza de la cartera. Los modelos 5 y 6 son similares, ambos maximizan entropía, pero mientras el modelo 5 controla la varianza con intervalos difusos, el modelo 6 solo minimiza la varianza. Finalmente, el modelo 7 con tres funciones objetivo, maximizando la entropía mientras controla rendimiento promedio y la varianza con medidas de entropía difusa.

3.2.1. Modelo 1: Modelo de Markowitz

Considerando $i \in \{1, ..., n\}$ activos, con retornos r_i. La variable de decisión x_i representa el peso relativo invertido en el portafolio, de acuerdo con la Ecuación (3.7). El objetivo es encontrar el portafolio óptimo a través de los pesos, cumpliendo con el retorno esperado en la Ecuación (3.8) y un límite de varianza en la Ecuación (3.9) del portafolio. Podemos construir el modelo base de Markowitz a partir de esta configuración inicial y las variantes y especializaciones que proponemos para su análisis. En la Ecuación (3.9), donde σ_i son las desviaciones estándar de los rendimientos de los activos y ρ_{ij} son sus coeficientes de correlación.

$$\sum_{i=1}^{n} x_i = 1 \tag{3.7}$$

$$r_p(\mathbf{x}) = \sum_{i=1}^{n} x_i \mathbb{E}\{r_i\} \tag{3.8}$$

$$\mathbb{V}(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j \sigma_i \sigma_j \rho_{ij} \tag{3.9}$$

Además, se define en la Ecuación (3.10) el modelo base de Markowitz (modelo 1), que minimiza la varianza y cumple con R_p para el retorno esperado del portafolio.

$$\begin{array}{lll} \underset{\{\mathbf{x}>0\}}{\text{mín}} & \mathbb{V}(\mathbf{x}) & \\ \text{sujeto a} & r_p(\mathbf{x}) & = R_p \\ & \sum_{i=1}^{n} x_i & = 1 \end{array} \tag{3.10}$$

Finalmente, es posible notar que en la Ecuación (3.10), la restricción $r_p(\mathbf{x}) = R_p$ también podría ser mayor o igual ($\geq$) como en [80].

3.2.2. Modelo 2: Maximización de la entropía de Shannon

El modelo 2 incluye la maximización de la entropía $H(\mathbf{x})$. En la Ecuación (3.11), se aplica este concepto al portafolio en el modelo base, lo cual ayuda a producir portafolios más diversificados. En este problema de optimización, los retornos esperados R_p y la varianza σ_p^2 están dentro de las restricciones. En la Ecuación (3.12), definimos el modelo 2 que corresponde a este caso; este modelo también fue considerado por al menos [15] y [81].

$$H(\mathbf{x}) = -\sum_{i=1}^{n} x_i \ln(x_i) \tag{3.11}$$

$$\begin{array}{lll} \underset{\{\mathbf{x}>0\}}{\text{máx}} & H(\mathbf{x}) & \\ \text{sujeto a} & r_p(\mathbf{x}) & = R_p \\ & \mathbb{V}(\mathbf{x}) & \leq \sigma_p^2 \\ & \sum_{i=1}^{n} x_i & = 1 \end{array} \tag{3.12}$$

3.2.3. Modelo 3: Maximización de la entropía de Shannon y mínima varianza

En la Ecuación (3.13), se muestra el modelo 3, que incluye la maximización de la entropía y la minimización de la varianza del portafolio. El modelo de dicha ecuación equilibra la función objetivo, otorgando igual importancia tanto a la maximización de la entropía como a la minimización de la varianza del portafolio resultante, mientras se requiere un nivel específico R_p para el retorno del portafolio.

$$\begin{array}{lll} \underset{\{\mathbf{x}>0\}}{\text{mín}} & \{V(\mathbf{x}),\ -H(\mathbf{x})\} & \\ \text{sujeto a} & r_p(\mathbf{x}) & = R_p \\ & \sum_{i=1}^{n} x_i & = 1 \end{array} \tag{3.13}$$

Modelo 4: Multicriterio con maximización de la entropía de Shannon y parámetros difusos en el retorno objetivo y la varianza

En el modelo 4, se introduce un esquema de optimización multicriterio, particularmente agregando parámetros difusos al retorno esperado objetivo y a la varianza del portafolio. De esta manera, este caso se define como el modelo biobjetivo o modelo 4 en la Ecuación (3.14), donde maximizamos la entropía y controlamos el retorno esperado y la varianza con intervalos trapezoidales.

$$\begin{array}{lllll} \underset{\{\mathbf{x}>0,\lambda\in[0,1]\}}{\text{máx}} & \{H(\mathbf{x}),\lambda\} & & & \\ \text{sujeto a} & r_p(\mathbf{x}) & \geq & \lambda r_1 + (1-\lambda)\,\underline{r} \\ & r_p(\mathbf{x}) & \leq & \lambda r_2 + (1-\lambda)\,\overline{r} \\ & \mathbb{V}(\mathbf{x}) & \geq & \lambda\sigma_1^2 + (1-\lambda)\,\underline{\sigma}^2 \\ & \mathbb{V}(\mathbf{x}) & \leq & \lambda\sigma_2^2 + (1-\lambda)\,\overline{\sigma}^2 \\ & \sum_{i=1}^{n} x_i & = & 1 \end{array} \tag{3.14}$$

En dicha ecuación, $(\underline{r}, r_1, r_2, \overline{r})$ y $(\underline{\sigma}^2, \sigma_1^2, \sigma_2^2, \overline{\sigma}^2)$ son los parámetros de los conjuntos difusos trapezoidales asociados con el retorno esperado y la varianza del portafolio. Para esto, se utiliza la forma trapezoidal de un parámetro difuso; en la Figura 3.1, mostrando cómo el retorno R_p está representado por un trapezoide y una función de pertenencia $\mu(R_p)$, que en la Ecuación (3.14), es controlada por la variable de decisión μ. Cabe señalar que el caso de la función de pertenencia trapezoidal para la varianza es análogo, y cuando $r_1 = r_2$ la forma del trapezoide se convierte en una función de pertenencia triangular.

Para el trapezoide del retorno, en cada momento, se calcula el retorno máximo posible y se define cada parámetro del trapezoide en función de este valor. El valor 0 siempre fue el límite inferior para el trapezoide del retorno porque nadie querría invertir en un portafolio con un retorno esperado negativo. De esta manera, se calculan los valores de volatilidad mínima y máxima para el caso de la volatilidad en cada momento. Así, se establece el parámetro del trapezoide, donde para el parámetro medio menor, se suma a σ mínimo el largo del intervalo de volatilidad realizada multiplicado por 0.25 y para el parámetro medio mayor a σ máximo se le resta el largo del intervalo multiplicado por

0.25.

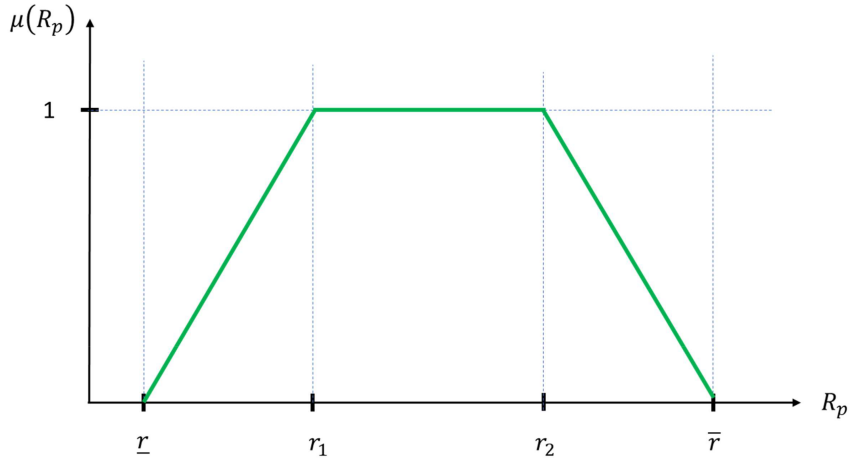

Figura 3.1: Figura explicativa para el parámetro difuso trapezoidal R_p.

3.2.4. Modelo 5: Multicriterio Shannon, varianza difusa y entropía difusa en el retorno objetivo

En la Ecuación (3.16), se define un modelo con un límite difuso trapezoidal. Este enfoque es triobjetivo, siendo su primera función objetivo es maximizar la entropía $H(\mathbf{x})$. Luego, el segundo objetivo es un enfoque de entropía difusa para maximizar la entropía del retorno difuso del portafolio $g_1(\alpha)$ (ver Ecuación (3.15)). Finalmente, el tercero es maximizar la pertenencia de la varianza usando la variable de decisión λ.

$$g_1(\alpha) = -\alpha \ln(\alpha) - (1-\alpha)\ln(1-\alpha) \tag{3.15}$$

$$\begin{aligned} \max_{\{\mathbf{x}>0,\alpha\in[0,1],\lambda\in[0,1]\}} \quad & \{H(\mathbf{x}), g_1(\alpha), \lambda\} & & \\ \text{sujeto a} \quad & r_p(\mathbf{x}) & = & \alpha\underline{r} + (1-\alpha)\overline{r} \\ & \mathbb{V}(\mathbf{x}) & \geq & \lambda\sigma_1^2 + (1-\lambda)\underline{\sigma}^2 \\ & \mathbb{V}(\mathbf{x}) & \leq & \lambda\sigma_2^2 + (1-\lambda)\overline{\sigma}^2 \\ & \sum_{i=1}^{n} x_i & = & 1 \end{aligned} \tag{3.16}$$

El modelo en la expresión de la Ecuación (3.17) es un caso particular de la Ecuación (3.16) en el cual $\sigma_p^2 = \sigma_1^2 = \sigma_2^2$ y $\lambda = 1$. En esta se representa la restricción asociada a la varianza como un límite superior, considerando que si existe alguna solución con una varianza menor que σ_p^2, entonces se considera como una buena solución.

$$\begin{array}{lll} \underset{\{\mathbf{x}>0,\alpha\in[0,1]\}}{\text{máx}} & \{H(\mathbf{x}), g_1(\alpha)\} & \\ \text{sujeto a} & r_p(\mathbf{x}) = \alpha\underline{r} + (1-\alpha)\overline{r} & \\ & \mathbb{V}(\mathbf{x}) \leq \sigma_p^2 & \\ & \sum_{i=1}^{n} x_i = 1 & \end{array} \tag{3.17}$$

De esta manera, se muestra la función de entropía difusa $g_1(\alpha)$ en la Figura 3.2. En este caso, α controla el valor de $R_p \in [\underline{r}, \overline{r}]$.

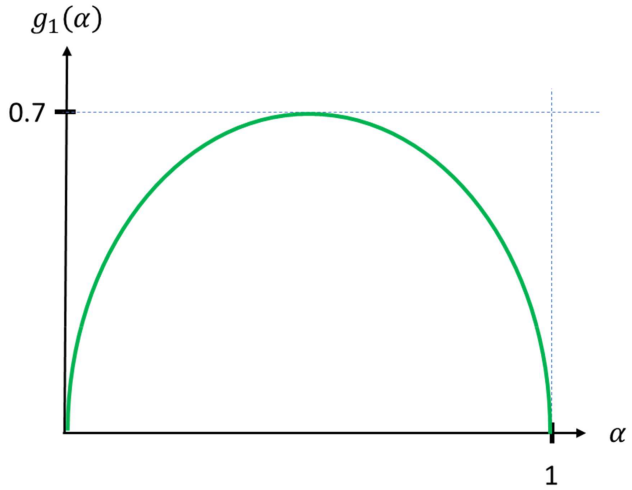

Figura 3.2: Figura explicativa para la función de entropía difusa $g_1(\alpha)$.

3.2.5. Modelo 6: Multicriterio Shannon, entropía difusa en el retorno objetivo y mínima varianza

En la Ecuación (3.18), se define el modelo 6 con una representación de entropía difusa para el retorno esperado y con tres funciones objetivo. En este contexto, la primera es la maximización de la entropía del portafolio, la segunda es la maximización de la entropía difusa del retorno esperado y la tercera es la minimización de la varianza del portafolio.

$$\begin{array}{lll} \underset{\{\mathbf{x}>0,\alpha\in[0,1]\}}{\text{máx}} & \{H(\mathbf{x}), g_1(\alpha), -\mathbb{V}(\mathbf{x})\} & \\ \text{sujeto a} & r_p(\mathbf{x}) = \alpha\underline{r} + (1-\alpha)\overline{r} & \\ & \sum_{i=1}^{n} x_i = 1 & \end{array} \tag{3.18}$$

3.2.6. Modelo 7: Multicriterio Shannon y entropía difusa en el retorno objetivo y la varianza

En la Ecuación (3.20), se propone el modelo 7, el cual tiene tres funciones objetivo. La primera es la maximización de la entropía del portafolio $H(\mathbf{x})$, la segunda es la maximización de la función de entropía difusa del retorno del portafolio $g_1(\alpha)$, y la tercera es la minimización de la varianza del portafolio $g_2(\beta)$, definida en la Ecuación (3.19). Además, es importante notar que la forma de la función de entropía difusa $g_2(\mathbf{x})$ es la misma que $g_1(\mathbf{x})$.

$$g_2(\beta) = -\beta \ln(\beta) - (1-\beta)\ln(1-\beta) \tag{3.19}$$

$$\begin{array}{lll} \underset{\{\mathbf{x}>0,\alpha\in[0,1],\beta\in[0,1]\}}{\text{máx}} & \{H(\mathbf{x}), g_1(\alpha), g_2(\beta)\} & \\ \text{sujeto a} & r_p(\mathbf{x}) = \alpha\underline{r} + (1-\alpha)\overline{r} & \\ & \mathbb{V}(\mathbf{x}) = \beta\underline{\sigma}^2 + (1-\beta)\overline{\sigma}^2 & \\ & \sum_{i=1}^{n} x_i = 1 & \end{array} \tag{3.20}$$

3.3. Aprendizaje reforzado profundo multiagente

En el aprendizaje supervisado, un algoritmo se entrena con conjuntos de datos etiquetados. Esto significa que, cuando el algoritmo realiza una determinación sobre un fragmento de información, puede usar las etiquetas incluidas con los datos para comprobar si esa determinación es correcta. Mientras tanto, con el aprendizaje no supervisado, los algoritmos se entrenan con datos que no contienen etiquetas ni información que el algoritmo pueda usar para comprobar sus determinaciones. De esta manera, el sistema ordena y clasifica los datos en función de los patrones que reconoce por sí mismo.

En aprendizaje por refuerzo no se tiene una "etiqueta de salida", por lo que no es de tipo supervisado. Si bien dichos algoritmos aprenden por sí mismos, tampoco son de tipo no supervisado, en donde se intenta clasificar grupos considerando alguna distancia entre muestras en la Figura 3.3.

Figura 3.3: Tipos de aprendizaje.

Luego, cuando se refuerza un aprendizaje, un sistema resuelve las tareas mediante la técnica de ensayo y error para tomar una serie de decisiones en una secuencia y lograr un resultado previsto, incluso en un entorno que no es sencillo. Con el aprendizaje de refuerzo, el algoritmo no usa conjuntos de datos para hacer determinaciones, sino información que recopila de un entorno.

Por otra parte, un aprendizaje se vuelve profundo cuando la modelación del problema se realiza a través de un subgrupo del aprendizaje automático (machine learning, ML), donde redes neuronales artificiales aprenden de grandes cantidades de datos, de esta manera se basa en capas de las redes neuronales, entrenadas con grandes cantidades de datos, configurando las neuronas en la red neuronal.

Cuando se combinan técnicas de aprendizaje profundo y aprendizaje de refuerzo, se da lugar a un tipo de aprendizaje automático denominado aprendizaje por refuerzo profundo. El aprendizaje de refuerzo profundo utiliza la misma técnica de toma de decisiones por

ensayo y error, y la misma consecución de objetivos complejos que el aprendizaje de refuerzo, pero con la funcionalidad del aprendizaje profundo a través de redes neuronales para procesar y dar sentido a grandes cantidades de datos no estructurados.

En consolidación de estás áreas, el aprendizaje por refuerzo de multiagente es entonces un subcampo del aprendizaje por refuerzo que se centra en estudiar el comportamiento de múltiples agentes de aprendizaje que conviven en un entorno compartido. Dentro de este, cada agente está motivado por sus propias recompensas y realiza acciones para promover sus propios intereses.

3.3.1. Aprendizaje reforzado

El Aprendizaje por refuerzo (RL) es una rama del aprendizaje automático donde un agente inteligente aprende cómo actuar dentro de un entorno o ambiente, buscando maximizar las recompensas a largo plazo dadas por un intérprete en función de ciertos objetivos de rendimiento definidos previamente [30].

En la Figura 3.4 se muestra un esquema general de cómo funciona el aprendizaje por refuerzo. En la primera iteración, el agente realiza algunas acciones aleatorias dentro del entorno. Estos entornos aleatorios, incluyendo el comportamiento de los mercados financieros, por lo general, pueden ser modelados como Procesos de Decisión de Markov (PDM) [5].

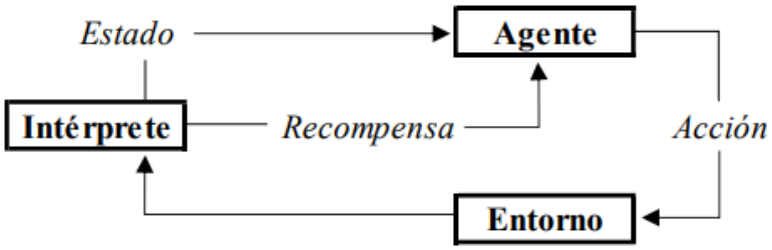

Figura 3.4: Diagrama de flujo de aprendizaje por refuerzo.

Los resultados de dicha iteración serán una observación del intérprete, quien, dependiendo de los objetivos de optimización definidos previamente de esta observación, otorgará una recompensa al agente por el buen o mal desempeño que obtuvo tras tomar esas acciones. El agente aprenderá de las observaciones y recompensas recibidas, y aplicará ese conocimiento en futuras iteraciones dentro del entorno, buscando maximizar la cantidad de recompensa que recibe.

Además, los primeros algoritmos de RL fueron entrenados para resolver problemas en entornos de baja dimensión [5]. Sin embargo, con la aparición de las redes neuronales

profundas, los algoritmos RL comienzan a ser más complejos, eficientes y útiles para resolver problemas en entornos más difíciles, dando cabida a los algoritmos de Aprendizaje por Refuerzo Profundo [82].

En el marco de la aparición de las redes neuronales profundas como solución para problemas de alta dimensionalidad, los estudios de aprendizaje por refuerzo se volcaron a la utilización de redes neuronales convolucionales como componentes de los agentes [82], y otras soluciones comenzaron a ser planteadas con el objetivo de lograr eficiencias necesarias en este tipo de entornos de grandes dimensiones.

A continuación, se definen algunos conceptos importantes comprender el capítulo metodológico de aprendizaje por refuerzo.

En primer lugar, los Procesos de Decisión de Markov (PDM) constituyen un marco común para resolver problemas de aprendizaje por refuerzo con algunos supuestos, como, por ejemplo, que el entorno es markoviano y observable [5] o parcialmente observable. Bajo esta premisa, el agente tendría que ser capaz de observar el entorno y luego, tomar decisiones dentro de este. Un algoritmo aprendizaje por refuerzo dentro de un PDM intenta encontrar las trayectorias para el agente dentro del entorno markoviano que maximizan la recompensa utilizando los siguientes parámetros [30]:

- Un estado,s, en el que se encuentra el agente, y que pertenece a un set de posibles estados S, siendo el estado inicial s_0.
- Una acción, a, que el agente toma en un determinado estado s, y que pertenece a un set de posibles acciones A.
- La recompensa inmediata, ρ, que el agente recibe por tomar una acción a en determinado estado s, llegando así a un estado nuevo, s'_t.
- Una política, π, que resulta de la distribución de probabilidad de tomar las acciones A encontrándose en determinado estado s.
- Una recompensa esperada, Q, de tomar acciones en un estado específico s y siguiendo una política π. Este concepto proviene del Q Learning [83].
- Una función de transición de estado,f, dada por la probabilidad de llegar al estado s' a partir del estado s por el hecho de tomar una acción a.
- Un factor de descuento, γ, que reduce el impacto de acciones futuras en el presente A.

En segundo lugar, se tiene la ecuación de Bellman y Q Learning, donde dado el número

de trayectorias, la política y los diferentes estados a los que se puede enfrentar el agente dentro del entorno, la recompensa esperada del agente por encontrarse en cierto estado, se plantea como una Función de Valor del Estado Bellman $V(s)$ [84], mejorada en Sutton [85], que a través de recursividad permite encontrar un valor de recompensa para el estado actual teniendo en cuenta los posibles estados futuros traídos a valor presente con el factor de descuento γ, tal como se muestra en la Ecuación (3.21).

$$V(s) = E_{\pi}[\rho_{t+1} + \gamma \cdot V(s_{t+1})/s_t = s] \tag{3.21}$$

En la misma línea, Watkins [86] posteriormente argumenta que el valor de la recompensa para el agente no debe obedecer únicamente al estado actual y estados futuros, sino que, además, debe depender de las acciones que el agente toma para llegar a diferentes estados, de manera que se plantea mapear todas estas posibles acciones basadas en la política que sigue el agente. Es entonces así, como la recompensa media ponderada de todas las posibles trayectorias individuales de una acción a partiendo de un estado s es la Función de Valor Estado-Acción $Q(a, s)$, siendo representada en la Ecuación (3.22) [83].

$$Q(a, s) = E_{\pi}[\rho_{t+1} + \gamma \cdot Q(a_{t+1}, s_{t+1})/a_t = a, s_t = s] \tag{3.22}$$

Esta función ayuda a mapear acciones futuras ligadas a un estado futuro relacionado, permitiendo así una convergencia más rápida y precisa en cantidad de instancias. Sin embargo, resulta ser un poco más exigente respecto a la capacidad de cómputo [5].

3.3.2. Métodos de aprendizaje reforzado profundo

Se extiende el marco tradicional de RL al incluir múltiples agentes que interactúan en un entorno compartido. En este contexto, cada agente tiene sus propios objetivos y puede aprender no solo de la interacción con el entorno, sino también de la interacción con otros agentes. Además, dentro del modelado del entorno financiero, es posible contextualizar una serie de elementos importantes a considerar:

- **Agentes**: En un entorno MARL para la gestión de portafolios, los agentes pueden representar diversos actores del mercado, como gestores de fondos o inversores individuales. Cada agente tiene un portafolio de activos y una estrategia de inversión que se adapta a lo largo del tiempo [87].

- **Entorno**: El entorno financiero incluye factores como la evolución de precios de los activos, cambios en las tasas de interés, noticias económicas, y la interacción entre agentes. Estos factores son observables para los agentes y condicionan sus

decisiones [52].

- **Dinámica de Interacción**: En términos de cooperación y competencia, los agentes en un entorno MARL pueden estar en competencia directa (como en la asignación de activos limitados) o en cooperación (como en estrategias de cobertura de riesgos). La interacción entre agentes afecta directamente la evolución de sus portafolios [88].

  Además, en algunos enfoques se puede dar un aprendizaje coordinado, los agentes pueden compartir información o coordinar estrategias para mejorar el rendimiento colectivo. Esto puede modelar consorcios de inversión o alianzas estratégicas [75] y será la base de la metodología MARL expuesta en esta tesis.

- **Optimización de Estrategias**: En términos de políticas de decisión, cada agente utiliza una política de decisión que es aprendida a través de episodios de interacción con el entorno. De esta manera, la política determina cómo se asignan los recursos entre los distintos activos en función del estado actual del mercado [89].

  Así mismo, esto se puede visualizar como exploración vs. explotación; los agentes deben equilibrar la exploración de nuevas estrategias de asignación de activos con la explotación de estrategias que han demostrado ser efectivas en el pasado [5].

- **Evaluación del Desempeño**: El éxito de un enfoque MARL en gestión de portafolios se evalúa mediante métricas como las ganancias, el rendimiento ajustado al riesgo o índice de Sharpe, la volatilidad del portafolio y la robustez frente a perturbaciones en el mercado [51]. Para esto, se utilizan datos históricos que permiten no solo entrenar los modelos sino validar la efectividad de las estrategias aprendidas, comparando el rendimiento de los agentes con estrategias paramétricas básicas de mercado y otros algoritmos de inversión [87].

  Además, la introducción de múltiples agentes y la naturaleza estocástica del entorno financiero aumentan significativamente la complejidad computacional. La optimización de políticas de decisión en tiempo real es un reto considerable [75]. El equilibrio entre agentes, entendido como garantizar que los agentes lleguen a un equilibrio estable donde ninguna estrategia dominante emerge de manera definitiva, es un desafío en sistemas multiagente.

  La convergencia hacia soluciones óptimas es un desafío [90], [5]. Del mismo modo, la generalización para desarrollar estrategias que se adapten a una amplia variedad de escenarios de mercado, incluyendo crisis financieras y cambios de régimen, es un área clave de la investigación multiagente realizada en esta tesis.

En términos de algoritmos, se destacan dos de una gran cantidad de variantes dis-

ponibles, debido a que serán utilizados como algoritmos de los modelos de aprendizaje planteados posteriormente en la metodología.

Algoritmo PPO (Proximal policy Optimization)

Combinación de algoritmos comunes de Aprendizaje por Refuerzo basados en políticas (sólo actor) con algoritmos basados en la función de valor del estado (sólo crítico). De este modo, los algoritmos actor-crítico como PPO optimizan utilizando la política, pero teniendo en cuenta la función de valor implícita del crítico. En consecuencia a [91], PPO utiliza un agente adicional, el crítico, que se encarga de revisar si la política es seguida por el agente que la ejecuta, el actor. Particularmente, en su arquitectura se entrega al agente crítico el resultado de la función de ventaja, ecuación (3.23), dada por la diferencia entre la función de valor estado-acción, y la función de valor del estado [88], lo que se interpreta como la ventaja de usar una política π sobre la función de valor del estado por sí sola. Para terminar, señalar que el algoritmo usa una política estocástica de tal manera que el actor genera una distribución de probabilidad sobre las posibles acciones y luego selecciona una acción de esa distribución (muy útil para la exploración).

$$A(a, s) = Q(a, s) - V(s) \tag{3.23}$$

Algoritmo DDPG (Deep Deterministic Policy Gradient)

Este algoritmo es una técnica de aprendizaje por refuerzo diseñada para problemas en los que las acciones están en un espacio continuo. En el contexto de trading, DDPG es útil para ajustar los pesos del portafolio de manera continua, permitiendo optimizar el rendimiento del portafolio a través del tiempo y con las siguientes consideraciones.

- **Acción Continua como Pesos del Portafolio**: Supongamos que el agente debe decidir cómo distribuir el capital entre distintos activos en un portafolio compuesto por N activos: $A_1, A_2, \ldots, A_N$. En cada momento, la acción continua que toma el agente se representa mediante un vector de pesos del portafolio:

$$\text{Acción} = [a_{A_1}, a_{A_2}, \ldots, a_{A_N}] \tag{3.24}$$

  Donde a_{A_i} representa el peso asignado al activo A_i. Estos pesos son valores continuos y suelen estar restringidos a un rango, como $[0, 1]$, para representar proporciones. Además, la suma de los pesos debe ser 1 para representar el total del capital invertido:

$$\sum_{i=1}^{N} a_{A_i} = 1 \tag{3.25}$$

- **Función de Recompensa** $J(\theta)$: El objetivo del agente en trading es maximizar la función de recompensa. La función $J(\theta)$ mide el rendimiento esperado a largo plazo y está dada por:

$$J(\theta) = \mathbb{E}_{s_0 \sim p(s_0)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \tag{3.26}$$

  Donde, θ representa los parámetros de la política, que genera las acciones continuas o pesos del portafolio en función del estado. Además, s_0 es el estado inicial, tomado de la distribución inicial $p(s_0)$. Por su parte, $\gamma \in [0, 1]$ es el factor de descuento, que controla la importancia de las recompensas futuras. Adicionalmente, $r(s_t, a_t)$ es la recompensa obtenida en el tiempo t, que representa los objetivos del portafolio con la asignación de pesos a_t.

- **Gradiente de Política**: Para optimizar el retorno $J(\theta)$, DDPG utiliza el gradiente de política, que permite ajustar los parámetros θ en función de la función de valor de acción $Q(s, a)$. Esto se expresa mediante la siguiente ecuación:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim \rho^\pi} \left[\nabla_\theta \pi_\theta(s) \nabla_a Q^\pi(s, a) \big|_{a=\pi_\theta(s)} \right] \tag{3.27}$$

  Dentro de esta última, $\pi_\theta(s)$ representa la política que produce el vector de pesos del portafolio (la acción) para el estado s. Además, $Q^\pi(s, a)$ corresponde a la función de valor de acción, que mide el rendimiento futuro esperado al tomar una acción a en un estado s y luego seguir la política π. Del mismo modo, ρ^π es la distribución de probabilidad de los estados cuando el agente sigue la política π.

Para implementar DDPG, algoritmo de actor-crítico, generalmente se usa una red neuronal que trabaja la función de política $\pi_\theta(s)$ y otra red que trabaja la función de valor de acción $Q(s, a)$. De esta manera, las redes son entrenadas en paralelo, donde una, actor, produce acciones basadas en los estados actuales, y otra, crítico, evalúa la calidad de ellas.

En términos comparativos, el algoritmo PPO debe usar muestras nuevas generadas por la red de manera reciente y DDPG usar muestras antiguas y recientes, una y otra vez para mejorar su rendimiento. Luego, PPO estará centrado en optimizar sus cálculos y DDPG en buscar la precisión.

Fundamentos Teórico conceptuales de modelación aplicados a los Agentes de AFPs.

La elección de arquitecturas y algoritmos responde a propiedades financieras del universo que gestiona cada agente y al hecho de que las decisiones de portafolio se formulan en un espacio de acciones continuo (pesos).

Renta Fija Nacional [RFN]: Los portafolios generados por este agente contienen principalmente activos de baja volatilidad, como bonos gubernamentales e instrumentos de alta calificación crediticia. Este agente prioriza la estabilidad sobre el rendimiento ajustado por riesgo, entonces, el modelo adecuado se basa en una red MLP y un algoritmo DDPG.

Renta Fija Internacional [RFI]: Este agente selecciona activos que ofrecen diversificación geográfica y reducen la correlación con los activos de renta fija nacional. Los portafolios presentan mayores pesos en bonos internacionales de mercados desarrollados y emergentes. En base a los objetivos del agente, la modelación queda basada en la integración de una red MLP y un algoritmo PPO.

Renta Variable Nacional [RVN]: Los activos seleccionados por este agente tienen altos índices de Sharpe, privilegiando acciones de alto crecimiento con buenos rendimientos históricos. Este agente genera portafolios más expuestos al riesgo en comparación con los otros agentes. En base a los objetivos del agente, es requerido un modelo basado en una red LSTM y un algoritmo DDPG.

Renta Variable Internacional [RVI]: Los activos seleccionados son aquellos que tienen altos indices de Sharpe, privilegiando acciones de alto crecimiento con buenos rendimientos históricos, priorizando mercados líquidos y acciones con alta rotación. Los portafolios generados son eficientes tanto en rendimiento ajustado al riesgo como en términos de costos operativos.La selección de portafolios queda modelada por una red CNN y un algoritmo DDPG.

Capítulo 4

Metodología

En esta sección, se describen los datos considerados para el modelo de optimización multicriterio y para el de aprendizaje reforzado multiagente.

Por un lado, en la optimización multicriterio y experimentación numérica se detalla cómo se ordenan las instancias numéricas descritas estadísticamente y por otro, se detallan los métodos de resolución de los problemas de optimización para finalmente, describir cómo se aplica TOPSIS para la toma de decisiones.

Respecto al modelo multiagente, primero se detallan estadísticamente los activos que gestiona cada agente, junto a los objetivos particulares que persigue cada uno.

Posteriormente, se presenta la estructura funcional de un modelo multiagente conformado por los agentes particulares, coordinados por el agente global, cuyo objetivo matemático es converger a una optimización conjunta de la cartera global, sin perder la convergencia y optimización de objetivos particulares.

Para finalizar, se detalla la construcción de un modelo híbrido a partir de la metodología multicriterio que considera la construcción del modelo 7 de entropía difusa expuesto en el capítulo 3, en la cual los portafolios de los agentes individuales obtenidos mediante los modelos particulares de RL son vistos como activos de una cartera a optimizar.

4.1. Descripción de los datos

4.1.1. Optimización multicriterio

La base de datos usada, consistente en una serie de precios diarios para cada activo y obtenida de la plataforma bloomberg, contiene registros mensuales de los precios de diferentes activos, convertidos en retornos diarios. Se hace uso de de los índices bursátiles de Estados Unidos (S & P 500), Europa (Euro Stoxx 50), Japón (Nikkei 225) y China (HSCEI), desde 2003 hasta 2023. La relevancia de utilizar índices en el portafolio radica en el hecho de que la estrategia es escalable a diferentes cantidades de inversión debido a que los volúmenes negociados de los instrumentos son muy altos.

Por un lado, el índice S&P 500, es uno de los índices bursátiles más importantes de Estados Unidos; se considera el índice más representativo de la situación real del mercado. Se basa en la capitalización de mercado de 500 grandes empresas, cuyas acciones cotizan en las bolsas NYSE o NASDAQ y captura aproximadamente el 80 % de toda la capitalización de mercado en Estados Unidos.

Por otro lado, el índice Euro Stoxx 50 representa el desempeño de las 50 empresas más grandes entre los 19 super-sectores en términos de capitalización de mercado en 11 países de la zona euro. Estos países son Alemania, Austria, Bélgica, España, Finlandia, Francia, Irlanda, Italia, Luxemburgo, Países Bajos y Portugal.

Además, el índice Nikkei 225, es el índice bursátil más popular en el mercado japonés, compuesto por los 225 valores más líquidos que cotizan en la Bolsa de Tokio.

Finalmente, el índice HSCEI es el principal índice bursátil chino de Hong Kong. Registra y supervisa los cambios diarios de las empresas más grandes de Hong Kong en el mercado bursátil. Está compuesto por 33 empresas que representan el 65 % de la Bolsa de Hong Kong.

4.1.2. Optimización bajo RL profundo multiagente

Se consideran los 4 segmentos que permiten a los inversionistas diversificar sus portafolios de acuerdo a su nivel de riesgo y horizonte de inversión como típicamente se realiza en una AFP, acorde al capítulo 3.1.2 . Estas áreas ofrecen una amplia gama de instrumentos en los segmentos de renta fija y variable, tanto a nivel nacional como internacional. De esta manera, permiten diversificar los portafolios y ajustar las estrategias individuales de acuerdo con el perfil de riesgo y objetivos. En el siguiente item se expresa la segmentación y los datos utilizados para la modelación de los agentes.

4.1.3. Descripción de las instancias numéricas

En términos de renta fija nacional se consideran bonos del gobierno y bonos corporativos con distintas duraciones. Para renta variable nacional se considera un portafolio compuesto por acciones de empresas chilenas que cotizan en la Bolsa de Comercio de Santiago y otras bolsas locales, para efectos de la modelación, se consideran algunas acciones relevantes dentro del mercado chileno y fondos invertidos de carteras de bancos nacionales. Para Renta Variable Internacional se incluye la compra de acciones de empresas que cotizan en bolsas extranjeras, como la Bolsa de Nueva York (NYSE) o la Bolsa de Londres (LSE), considerando para la modelación, el índice de bonos del tesoro americano y bonos de mercados emergentes internacionales, el índice global agregado, High Yield y High Grade. A continuación se detallan las carteras de cada agente conforme sus activos y objetivos.

Renta Fija Nacional

Para el análisis de los datos renta fija nacional, se utiliza una ventana de tiempo de 6 años de información semanal, permitiendo observar la evolución de precios y retornos de estos bonos en un horizonte de largo plazo. En este contexto, los datos considerados contienen series históricas de precios ajustados en CLP, desde el 31 de julio de 2018 hasta el 22 de octubre de 2024, cubriendo aproximadamente seis años de información diaria para los siguientes bonos, con características específicas en términos de duración, emisor y denominación: Bonos de Gobiernos en UF correspondientes a duraciones de 4,8 y 16 años y Bonos Corporativos en UF de Duración 5 años. La evolución de los bonos gubernamentales y corporativos a lo largo del tiempo se muestra en la Figura 4.1

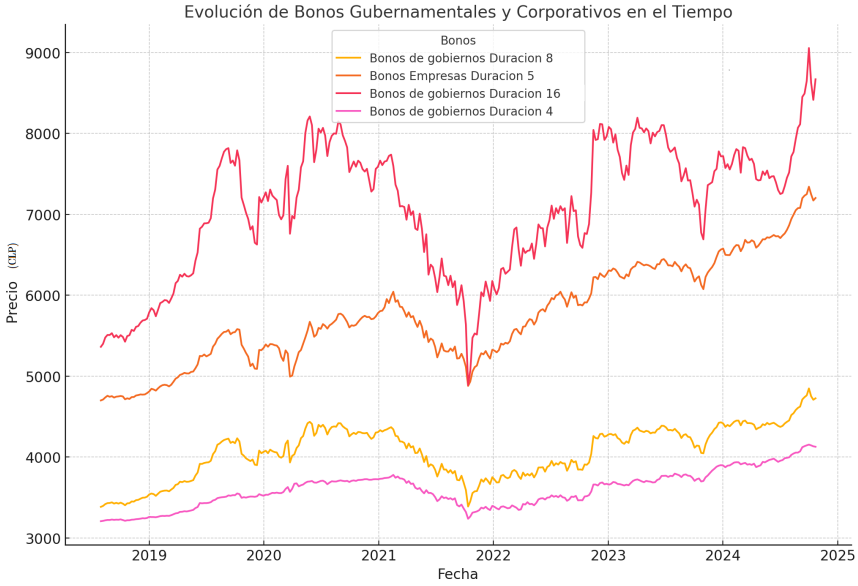

Figura 4.1: Evolución de Bonos Gubernamentales y Corporativos en el tiempo.

Del mismo modo, las estadísticas descriptivas de los precios de bonos en UF están dados por la siguiente tabla:

Cuadro 4.1: Estadísticas descriptivas de los precios de bonos en UF.

Tipo Bono	**Duración (años)**	**Promedio (UF)**	**Desv. Est. (UF)**	**Mín.-Máx. (UF)**
Gobierno	4	3595.26	223.99	3205.76 – 4152.55
	8	4059.75	325.00	3383.25 – 4846.88
	16	7061.10	818.94	4888.55 – 9054.59
Empresas	5	5750.78	624.92	4698.94 – 7340.39

A partir de la Figura 4.1 y la Tabla 4.1, es posible visualizar como los bonos con corta duración (4 años) muestran una trayectoria estable y con baja volatilidad a lo largo del tiempo. Esto refleja su menor sensibilidad a las fluctuaciones de las tasas de interés, ya que los plazos cortos minimizan el riesgo. En este contexto, es un instrumento atractivo para inversionistas conservadores que buscan estabilidad y menor riesgo.

Del mismo modo, los bonos de duración intermedia (8 años) presentan un comportamiento más dinámico que los de 4 años, con mayor volatilidad en ciertos períodos. Si bien mantienen estabilidad en el largo plazo, muestran mayor sensibilidad a cambios en las condiciones económicas y tasas de interés. En este contexto, son adecuados para inversionistas que buscan un equilibrio entre riesgo y retorno.

Además, los bonos de larga duración (16 años) registran los valores más altos en términos de precios, pero también muestran alta volatilidad. Durante períodos de incertidumbre económica o políticas monetarias restrictivas (como en 2021-2022), se observan caídas abruptas. Esta sensibilidad a las tasas de interés refleja el mayor riesgo asociado al largo plazo, pero también potenciales mayores rendimientos.

Los bonos corporativos (empresas 5 años) exhiben una trayectoria intermedia, con precios más altos que los bonos gubernamentales de 4 y 8 años. Sin embargo, muestran moderada volatilidad debido al riesgo crediticio inherente a los emisores corporativos. A pesar de esto, los rendimientos son atractivos y los episodios de caídas son menos pronunciados que en los bonos de larga duración.

Para este tipo de activos se desarrolla un modelo de optimización de portafolios utilizando aprendizaje por refuerzo profundo con un único agente y además un modelo de optimización de portafolios coordinado , en este último caso, el agente a través del modelo MARL, estará siendo supervisado por un coordinador en términos de mejorar la gestión global de la AFP.

Dentro de este modelo se entrena un agente de inteligencia artificial que buscará maximizar el rendimiento ajustado por riesgo del portafolio, manteniendo la estabilidad ante las fluctuaciones del mercado, sobre las siguientes métricas de mercado,a demás de la minimización inherente de los costos de movimientos en el portafolio:

- **Retornos**:Rendimiento para cada tipo de bono calculado a partir de los cambios de precios, permite al agente identificar activos con mejores proyecciones.
- **índice de Sharpe**: Indicador de la eficiencia, entendida como el retorno en función de la variabilidad de los mismos permitiendo medir el retonro ajustado riesgo inherente de cada bono.
- **Duración y Convexidad**: Medidas que representan la sensibilidad ante cambios en las tasas de interés, particularmente relevantes para bonos de mayor duración, como los de 16 años. En renta fija nacional, son métricas claves porque los flujos de efectivo están en moneda local.

Renta Fija Internacional

En términos de los datos renta fija internacional, nuevamente se considera el periodo comprendido entre el 31 de julio de 2018 y el 22 de octubre de 2024. Los datos incluyen los distintos índices relacionados como Bonos del Tesoro Americano, el índice global agregado (Global AGGREGATE), High Yield, High Grade y el índice corporativo de mercados emergentes (CEMBI).

Cabe destacar que los Bonos del Tesoro Americano muestran órdenes de magnitud significativamente superiores a los demás indicadores. Para abarcar esto, se cuenta con la Figura 4.2, la cual contiene dos gráficas insertas. La primera muestra la evolución de este indicador, que tiene valores significativamente más altos que los demás, mientras que la segunda contiene los índices Global AGGREGATE, High Yield, High Grade y CEMBI, que tienen una escala similar entre sí.

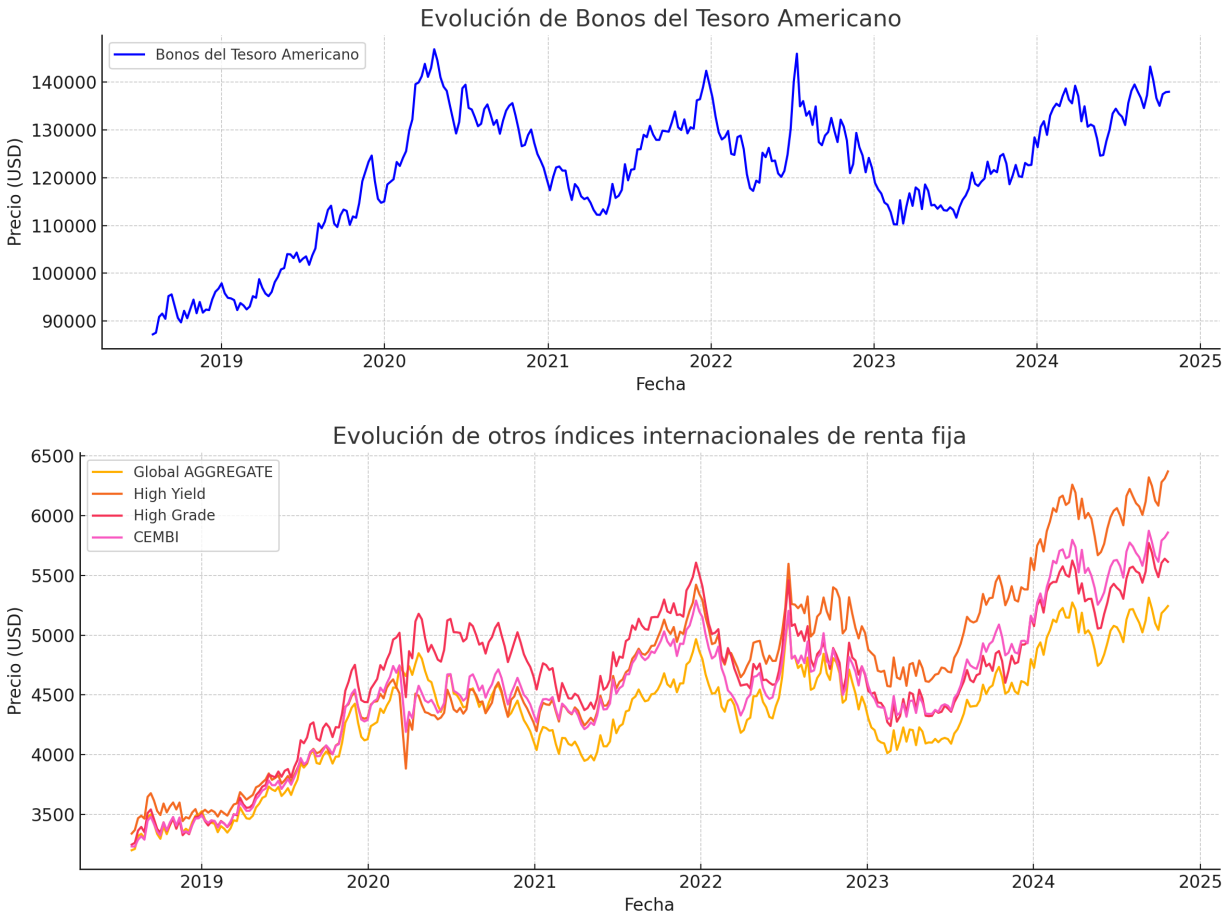

Figura 4.2: Evolución de activos de RFI en el tiempo.

Además, la siguiente tabla permite situarse estadisticamente en la cartera de inversiones de este agente, compuesta por los 5 índices.

Cuadro 4.2: Estadísticas descriptivas de los precios de activos de renta fija internacional.

Tipo de Bono	**Promedio (CLP)**	**Desv.Est (CLP)**	**Mín. (USD)**	**Máx. (CLP)**
Tesoro Americano	120920.98	13837.38	87209.37	146895.89
Global AGGREGATE	4330.70	493.14	3200.51	5313.89
High Yield	4708.79	736.68	3339.92	6370.24
High Grade	4640.46	601.40	3247.54	5771.27
CEMBI	4535.93	614.10	3230.87	5873.52

Los datos de Renta Fija Internacional (RFI) reflejan una combinación de activos con diferentes niveles de riesgo, retorno y sensibilidad a las tasas de interés y el riesgo país. Cada uno de estos activos cumple un rol clave dentro de una cartera diversificada, optimizando el equilibrio entre rentabilidad y estabilidad.

Además, A partir de la Figura 4.2 y la Tabla 4.2, los bonos del Tesoro Americano se presentan con el mayor promedio de precios (120920 CLP). Estos bonos representan la opción más segura dentro del conjunto, sirviendo como activo refugio en momentos de crisis. De esta manera, su presencia en una cartera ayuda a reducir la volatilidad y proveer estabilidad en periodos de incertidumbre económica.

Respecto al Global Aggregate, su precio promedio más bajo (4330 CLP) y una desviación estándar relativamente baja indican que este índice es una opción diversificada en sí misma, capturando el comportamiento del mercado global de renta fija. Su incorporación aporta diversificación internacional y exposición a distintas economías con menor riesgo que la renta variable. Del mismo modo, el índice High Yield presenta un precio medio de 4708 CLP y una mayor volatilidad (736 CLP), lo que refleja su mayor riesgo crediticio en comparación con otros bonos de renta fija. Su inclusión en una cartera puede mejorar el rendimiento, pero requiere una gestión activa del riesgo de impago.

Además, el índice High Grade, con un promedio de 4640 CLP y menor desviación estándar (601 CLP) que los bonos High Yield, ofrece un punto intermedio entre rentabilidad y seguridad. De esta manera, es una opción ideal para inversionistas que buscan estabilidad con rendimientos moderados en mercados desarrollados.

Finalmente, CEMBI (Bonos Corporativos de Mercados Emergentes) presenta un precio promedio de 4535 CLP, dado que refleja por definición el comportamiento de la deuda corporativa en economías emergentes. Si bien su rentabilidad puede ser atractiva, es importante que su riesgo país y exposición a monedas extranjeras sean gestionados adecuadamente.

Cada activo considerado dentro de la cartera de renta fija internacional contribuye de manera distinta a la optimización de la cartera, de manera esquemática:

- **Bonos del Tesoro Americano**: Estabilidad y refugio en momentos de crisis.
- **Global Aggregate**: Diversificación global con menor volatilidad.
- **High Yield**: Mayor retorno con riesgo crediticio elevado.
- **High Grade**: Compromiso entre seguridad y rendimiento.
- **CEMBI**: Exposición a mercados emergentes con potencial de crecimiento.

Incluir una combinación de estos activos en una cartera permite maximizar el rendimiento ajustado al riesgo, optimizando la relación entre rentabilidad y volatilidad. Su comportamiento varía según el ciclo económico, por lo que una gestión activa es clave

para balancear estabilidad y retorno esperados. En este contexto, matemáticamente, se hace referencia a un agente que sea capaz de captar distintos escenarios y adaptarse a ellos respondiendo de manera eficiente y rentable en escenarios complejos que pueden no ser representativos de eventos pasados.

Para este tipo de activos se entrena otro agente neuronal que buscará maximizar el retorno, el rendimiento ajustado por riesgo y minimizar los costos de movimientos del portafolio,, siendo parte del modelo de aprendizaje por refuerzo profundo a nivel uniagente y a nivel MARL como segundo agente involucrado.

Como antecedente comparativo con RFN, en RFI las variaciones en los tipos de cambio pueden afectar más el rendimiento total que la duración o convexidad del bono. La depreciación o apreciación de la moneda local frente a la divisa del bono puede tener un impacto significativo en los retornos. Por otra parte, los bonos internacionales están sujetos a decisiones de bancos centrales foráneos, lo que introduce factores exógenos difíciles de anticipar con métricas de sensibilidad como la duración. Factores como riesgo soberano, diferencial de tasas de interés y spreads de crédito internacionales pueden ser más determinantes que la duración y la convexidad. En muchas instituciones que invierten en renta fija internacional emplean coberturas en derivados (swaps de tasas o forwards de divisas), lo que altera el impacto real de la duración y la convexidad. Por lo tanto, no son variables a utilizar en la función de recompensa del modelo que gestiona RFI.

Renta Variable Nacional

Los datos de renta variable nacional, corresponden a una serie de precios diarios de un total de 10 distintos activos financieros del mercado, listados a continuación. Los datos cubren nuevamente el período comprendido entre el 31 de julio de 2018 y el 22 de octubre de 2024, con información diaria. De esta manera, contiene una cantidad de observaciones suficiente para permitir analizar la evolución de estos activos a lo largo del tiempo, fluctuaciones y períodos de estabilidad o volatilidad.

- **SQM**: Precio de la acción de Sociedad Química y Minera de Chile.
- **CHILE**: Índice IPSA, que representa el mercado accionario chileno.
- **BSANTANDER**: Banco Santander Chile, relevante en el sector financiero.
- **FALABELLA**: Empresa líder en retail en América Latina.
- **CENCOSUD**: Importante mayorista en Chile.
- **BCI**: Banco de Crédito e Inversiones, clave en el sistema financiero chileno.

- **COPEC**: Compañía de Petróleos de Chile, líder en combustibles y energía.
- **LTM**: Latam Airlines, principal aerolínea en América Latina.
- **ENELAM**: Enel Américas, relevante en el sector energético.
- **CMPC**: Compañía Manufacturera de Papeles y Cartones.

Dentro de la muestra se incluyen compañías representativas de distintos sectores económicos, lo que otorga diversidad al análisis. Entre los activos considerados se encuentran SQM, una de las principales empresas del sector minero y químico en Chile, junto con entidades financieras como Banco de Chile y Banco Santander. Del mismo modo, se incluyen grandes firmas del sector retail, tales como Falabella y Cencosud, además de compañías del sector energético y forestal como Enel Américas, Copec y CMPC. Finalmente, el listado se completa con BCI, junto a Latam Airlines (LTM), empresa clave en la industria del transporte aéreo en la región. A nivel estadístico, uno de los aspectos más relevantes del conjunto de datos es la variabilidad en el comportamiento de los distintos activos. En este contexto, cabe señalar a LTM, con unos peaks extremadamente fuera de los rangos estadísticos y empresas con una exposición más fuerte a ciclos económicos y fluctuaciones internacionales, como SQM, presentan una volatilidad significativamente mayor en comparación con compañías del sector bancario, como Banco de Chile y Banco Santander, cuyos precios han mostrado mayor estabilidad a lo largo del período.

A continuación, se muestra una gráfica que permite visualizar la evolución de retornos de RVN en el tiempo:

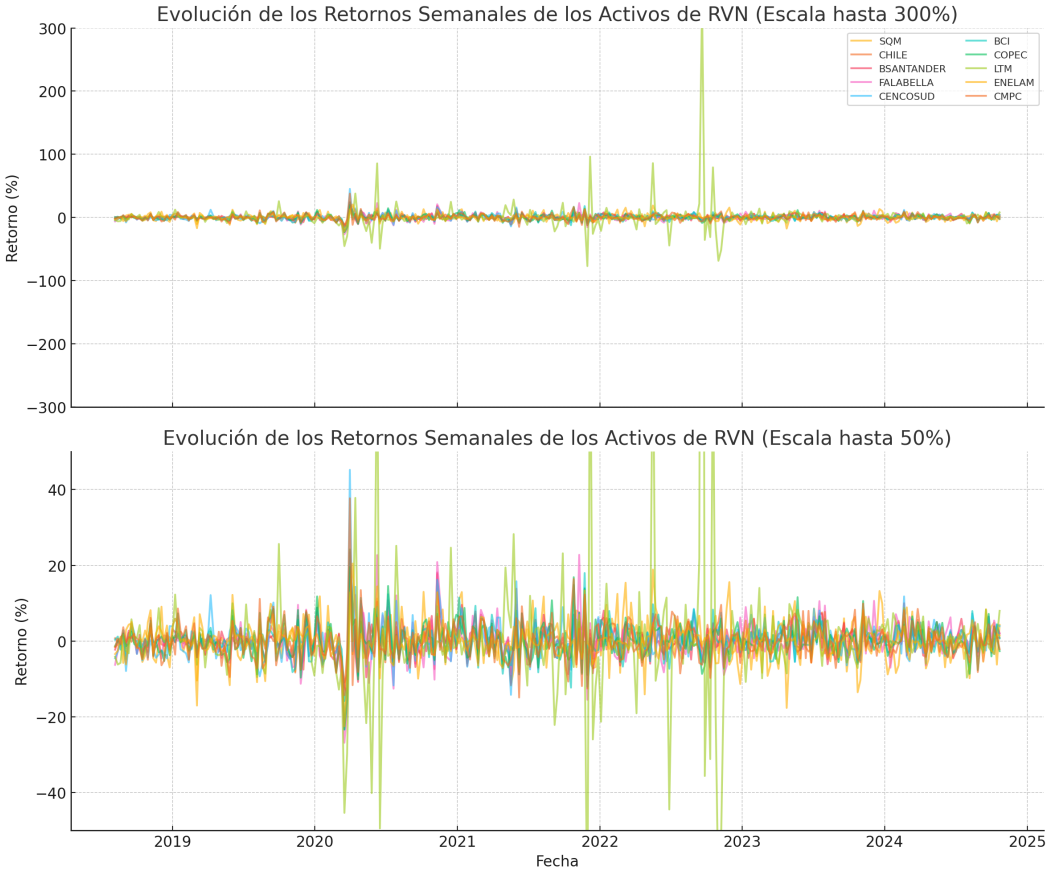

Figura 4.3: Evolución de retornos de RVN en el tiempo.

Además, se resume en la siguiente tabla las estadísticas de precios para la renta variable nacional:

Cuadro 4.3: Estadísticas de Precios de RVN (CLP).

Activo	**Promedio (CLP)**	**Desv. Est. (CLP)**	**Mínimo (CLP)**	**Máximo (CLP)**
SQM	37418.49	19423.33	12331.95	86897.93
CHILE	72.14	16.88	43.46	117.20
BSANTANDER	35.23	5.98	20.62	48.38
FALABELLA	2847.59	971.64	1441.82	5505.12
CENCOSUD	1156.30	349.56	420.14	1975.00
BCI	24878.55	2339.56	19650.00	30560.00
COPEC	8152.24	754.29	6400.00	9650.00
LTM	385.17	140.89	200.00	690.00
ENELAM	88.33	12.15	60.00	115.00
CMPC	1791.02	295.34	1350.00	2450.00

Del mismo modo, en este acápite se expresan las estadísticas de los retornos funda-

mentado en las características, propias de la inestabilidad y pagos de este mercado.

Cuadro 4.4: Estadísticas de Retornos de RVN (%).

Tipo de Activo	**Promedio (%)**	**Desv. Est. (%)**	**Mínimo (%)**	**Máximo (%)**
SQM	0.33	5.87	-22.84	24.06
CHILE	0.22	3.70	-11.04	21.21
BSANTANDER	0.15	4.10	-13.85	23.36
FALABELLA	0.01	5.43	-26.85	23.27
CENCOSUD	0.29	5.21	-23.41	45.20
BCI	0.14	4.38	-15.83	24.20
COPEC	0.02	4.61	-23.32	19.51
LTM	-0.76	24.12	-77.07	343.78
ENELAM	0.07	4.02	-22.52	20.46
CMPC	0.08	5.14	-14.91	37.67

A partir de la Figura 4.3 y la Tabla 4.4, es posible notar que los activos con mayor rentabilidad en el período analizado incluyen SQM, Cencosud y COPEC. Estos sectores presentan características estructurales que explican su desempeño, por un lado, SQM es una de las principales compañías productoras de litio y otros minerales esenciales para la industria global. La creciente demanda de litio para baterías ha impulsado su rentabilidad, reflejándose en un retorno promedio superior a otros activos en el mercado nacional. Sin embargo, también es un sector con alta exposición a precios internacionales de commodities, lo que genera volatilidad. Por otro lado, Cencosud ha mostrado un crecimiento sostenido a pesar de la incertidumbre económica, logrando mantener retornos competitivos. Su rendimiento ha sido impulsado por estrategias de expansión y consolidación en mercados clave. Además, la rentabilidad de COPEC se ve favorecida por su integración en el sector energético y de combustibles; sus ingresos están altamente correlacionados con la demanda interna de combustibles y el precio del petróleo. A pesar de que este sector puede enfrentar variaciones en márgenes debido a fluctuaciones del precio del crudo, la estabilidad en el consumo energético ha permitido retornos consistentes.

Los activos que han experimentado mayor volatilidad incluyen SQM, Falabella y LTM, siendo explicada por la sensibilidad a factores macroeconómicos y específicos de cada industria. En este contexto, si bien SQM es uno de los activos más rentables, también muestra una elevada volatilidad debido a la fuerte dependencia del precio del litio en los mercados internacionales. Falabella, a diferencia de Cencosud, ha mostrado una volatilidad superior debido a su exposición a las incertidumbres económicas, junto con el crecimiento del comercio electrónico, generando fluctuaciones importantes en su rentabilidad, afectando su estabilidad en los mercados bursátiles. La volatilidad de LA ha sido una de las más

elevadas en el período analizado. Esta industria es altamente cíclica y dependiente del precio del combustible, la demanda de pasajeros y los factores macroeconómicos. De forma adicional, es importante notar que la pandemia de COVID-19 y sus efectos en la recuperación del tráfico aéreo han impactado directamente en su estabilidad financiera.

Como análisis final, se evidencia que los activos con mayor rentabilidad no necesariamente son los más estables, y la relación entre rentabilidad y riesgo varía según la industria. Mientras que sectores como minería y energía presentan retornos atractivos, también están sujetos a alta volatilidad debido a la dependencia de factores externos. Por otro lado, empresas de retail y transporte han enfrentado fluctuaciones considerables, impactadas por cambios estructurales en sus mercados y por la incertidumbre económica.

Este tipo de análisis es crucial para la toma de decisiones de inversión, ya que permite a los inversionistas equilibrar rentabilidad esperada con tolerancia al riesgo, considerando la naturaleza de cada industria y su comportamiento en particular en este mercado. Finalmente, ENELAM y SQM son los activos más relevantes para inversores interesados en alta volatilidad y oportunidades globales. Por otro lado, BSANTANDER, CENCOSUD, y CMPC representan opciones más estables y predecibles. Mientras tanto, los índices CHILE y ENELAM reflejan el comportamiento del mercado nacional y sectores específicos, proporcionando una visión más amplia del entorno económico chileno.

Renta Variable internacional

En renta variable internacional (RVI), los factores de riesgo y rendimiento se relacionan más directamente con el comportamiento de los mercados bursátiles globales y su interacción con variables macroeconómicas, eventos geopolíticos y ciclos económicos regionales. A diferencia de renta fija, donde las tasas de interés y los flujos de caja son fundamentales, la renta variable internacional presenta una dinámica más compleja, altamente correlacionada con crecimiento económico, innovación tecnológica y política monetaria global.

En este contexto, el conjunto de datos correspondiente a RVI contempla precios semanales de siete índices o activos representativos de los principales mercados accionarios globales: Estados Unidos, Europa, Japón, mercados emergentes, Asia (excluyendo Japón), China e India . Esta base cubre el mismo período comprendido entre el 31 de julio de 2018 y el 22 de octubre de 2024. Cada uno de estos activos representa una región económica con dinámicas propias, por lo cual su inclusión en un portafolio aporta diversificación geográfica, sectorial y política. A continuación, se describen estos componentes:

- **SP 500**: Índice bursátil que representa el mercado accionario estadounidense. Altamente líquido y referencia global en renta variable, con fuerte exposición a sectores tecnológicos, financieros e industriales.

- **EUROPA**: Conjunto representativo del mercado accionario europeo. Sensible a decisiones del Banco Central Europeo ya la cohesión económica del bloque. Diversificado en consumo, energía y fabricación.
- **JAPÓN**: Índice del mercado accionario japonés, asociado a grandes conglomerados industriales y tecnológicos. Mercado maduro con comportamiento menos correlacionado a EE.UU.
- **EMERGENTES**: Índice que agrupa mercados en desarrollo (Latinoamérica, Europa del Este, Asia). Activo más volátil del conjunto, sensible a flujos de capital y precios de commodities.
- **ASIA**: Compuesto por economías asiáticas excluyendo Japón. Alta exposición a fabricación, tecnología y crecimiento económico. Correlación media con EE.UU. y volatilidad moderada.
- **CHINA**: Índice representativo del mercado accionario chino. Elevada volatilidad, regulación impredecible, pero importancia estratégica creciente a nivel mundial.
- **INDIA**: Índice del mercado indio, caracterizado por alto crecimiento y políticas internas expansivas. Su comportamiento combina rasgos de mercados emergentes con fundamentos macroeconómicos sólidos.

Cuadro 4.5: Estadísticas de Precios de RVI.

Activo	**Promedio (CLP)**	**Desv. Est. (CLP)**	**Mínimo (CLP)**	**Máximo (CLP)**
SP 500	326,116.04	106,410.97	158,540.13	587,432.11
EUROPA	121,439.09	28,608.46	77,345.59	187,551.30
JAPON	134,434.96	26,451.04	91,459.10	200,434.11
EMERGENTES	29,080.41	4,622.27	20,591.37	39,250.25
ASIA	131,038.77	22,842.24	87,957.93	180,490.52
CHINA	3,863.98	808.54	2,211.67	5,597.74
INDIA	5,518.49	1,818.88	2,870.84	9,872.82

Cuadro 4.6: Estadísticas de retornos semanales de RVI.

Activo	**Promedio (%)**	**Desv. Est. (%)**	**Mínimo (%)**	**Máximo (%)**
SP 500	0.41	2.62	-10.55	8.46
EUROPA	0.27	2.82	-15.84	11.52
JAPON	0.23	2.56	-11.88	10.22
EMERGENTES	0.21	2.44	-12.66	7.74
ASIA	0.23	2.61	-11.54	8.59
CHINA	0.25	3.33	-10.13	18.66
INDIA	0.36	3.03	-13.89	11.30

En términos estadísticos,de acuerdo a la Tabla 4.4 se observa una clara diferenciación entre los activos más estables y aquellos más volátiles. El índice S&P 500 presenta una media elevada y un coeficiente de variación moderado (0.33), lo que evidencia su relevancia como activo ancla en el portafolio global. Europa y Japón muestran niveles promedio de precios intermedios y una dispersión más controlada, con coeficientes de variación entre 0,20 y 0,24, lo que refleja su comportamiento más predecible en comparación con economías emergentes. Los activos con menor media y mayor desviación estándar son los de mercados emergentes y Asia , lo que muestra tanto su volatilidad controlada como su menor peso relativo. Sin embargo, su valor dentro de la cartera no radica tanto en su nivel de precios, sino en su correlación baja con activos desarrollados, lo que entrega beneficios en términos de diversificación. Los activos de China e India, forman parte fundamental del análisis al ser economías en expansión, con capacidad de liderazgo global en sectores como tecnología, manufactura, servicios digitales y energía. Su volatilidad se encuentra en el rango más alto de los activos de la cartera, pero el rendimiento ajustado por riesgo en horizontes de largo plazo es atractivo.

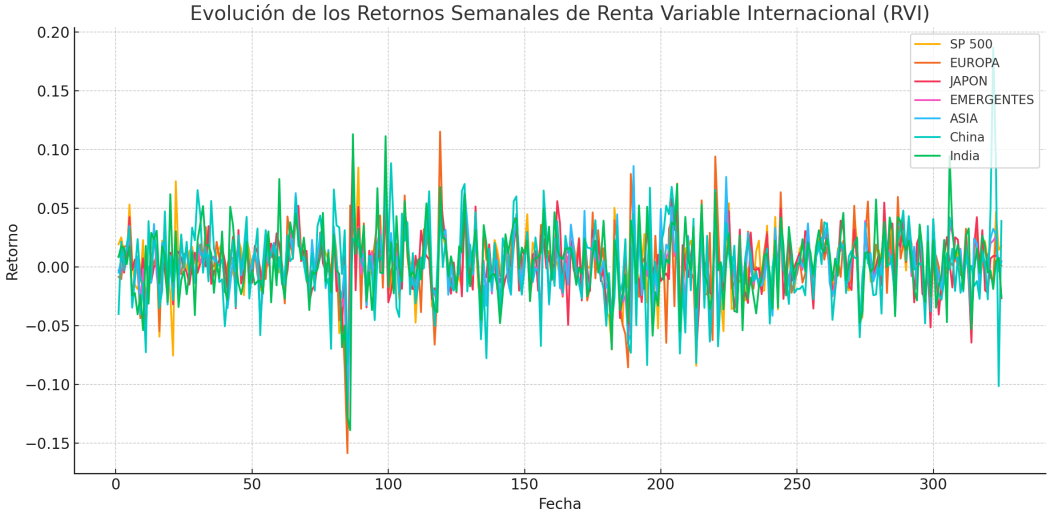

Figura 4.4: Evolución de los Retornos de Variable Internacional

.

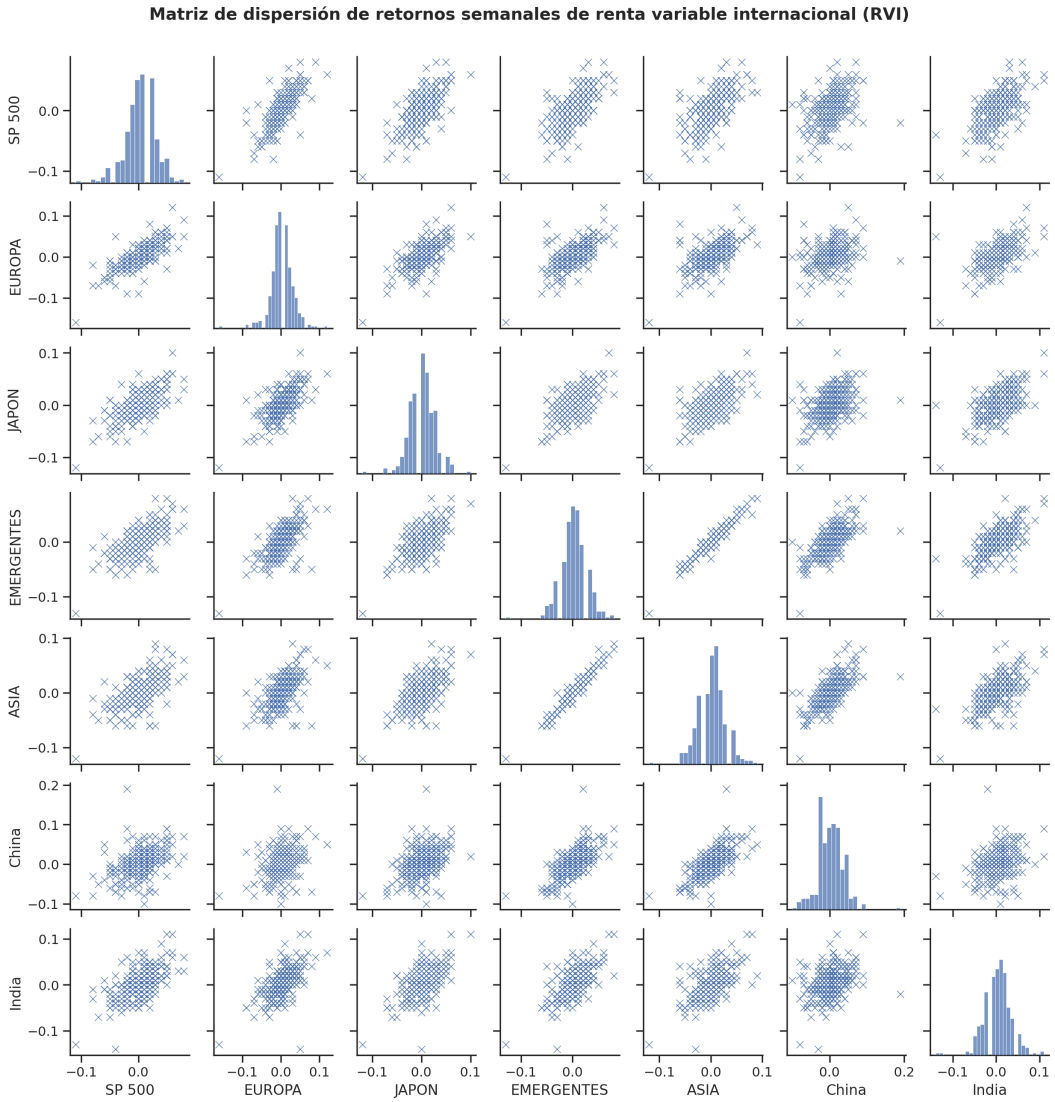

Figura 4.5: Matriz de dispersión de retornos.

Un análisis exploratorio de Figura 4.4, la Tabla 4.6 y la Figura 4.5, correspondiente a la matriz de dispersión de los retornos del portafolio presenta variaciones en los diferentes índices bursátiles internacionales que lo componen. Los aspectos más relevantes son:

- Los activos Emergentes, Europa y Japón presentan algunos picos más pronunciados, reflejando una mayor sensibilidad a eventos globales. En contraste, el S&P 500 que muestra una evolución más estable, la curva que describe es prácticamente imperceptible, con menor dispersión relativa.

- En gran medida los activos se mueven en la misma dirección, lo cual evidencia correlación positiva durante eventos sistémicos. No obstante, también se aprecian momentos de divergencia que reflejan oportunidades de diversificación internacional.

A pesar de evidenciar correlaciones durante eventos sistémicos, las diferencias en magnitud, volatilidad y dirección de los retornos refuerzan el valor de una estrategia de diversificación global. La renta variable internacional, en este contexto, entrega exposición a oportunidades diferenciadas, aunque exige una gestión consciente del riesgo sistemático y regional.

- Se identifican caídas abruptas simultáneas en todos los activos en momentos específicos, lo cual sugiere la presencia de eventos globales como crisis financieras, afectando de forma sincronizada.

- **E**l índice SP 500 destaca por mantener retornos positivos más estables, mientras que China, India y Asia presentan mayor dispersión con retornos en promedio más altos que los mercados Emergentes y Asia, lo cual indica la necesidad de estrategias activas de manejo del riesgo o coberturas.

Los activos incluidos en RVI permiten construir un portafolio globalmente diversificado, donde la exposición a economías desarrolladas (EE.UU., Europa, Japón) se combina con el dinamismo y volatilidad controlada de mercados emergentes y asiáticos. Este tipo de composición permite capturar oportunidades de crecimiento, mantener exposición a sectores líderes globales y, al mismo tiempo, distribuir el riesgo geográfico, sectorial y económico de forma eficiente. La volatilidad, lejos de ser una desventaja, se convierte en el precio a pagar por acceder a retornos potenciales superiores en una economía global en transición.

4.2. Optimización multicriterio

A continuación, se describen los datos considerados para la experimentación numérica. Además, se explica cómo se ordenan las instancias numéricas y se detalla cómo se resuelven los problemas de optimización, para finalmente describir cómo se aplica TOPSIS.

4.2.1. Descripción de las instancias numéricas

De todo el conjunto de datos, con valores mensuales desde el año 2003 hasta 2023, correspondientes a 235 muestras de cada índice bursátil, dividimos el conjunto de datos en cinco subconjuntos no superpuestos mediante una partición temporal, con la misma longitud (47 muestras). El objetivo fue probar el modelo bajo diferentes condiciones de mercado, denominadas conjuntos de análisis independientes ($S1, S2, S3, S4, S5$). A continuación, se muestra en la Tabla 4.7 el intervalo de tiempo para cada escenario.

Cuadro 4.7: Subconjuntos.

Fechas/ Escenario	1	2	3	4	5
Fecha de inicio	31-05-2003	31-05-2007	30-04-2011	31-03-2016	29-02-2020
Fecha de fin	31-03-2007	31-03-2011	28-02-2015	31-01-2020	31-12-2023

Para evaluar el desempeño de cada modelo en cada subconjunto, se utiliza un enfoque de ventana móvil, donde se calcula los rendimientos esperados y las matrices de covarianza para cada ventana. De esta manera, la metodología de ventana temporal móvil queda expresada en la Figura 4.6.

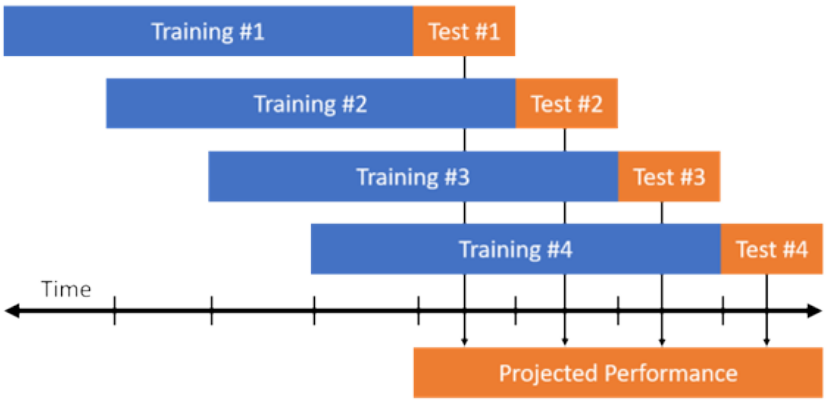

Figura 4.6: Ventana temporal móvil.

Además, se considera relevante explicar que, al utilizar cuatro índices, las varianzas y covarianzas no constituyen matrices de alta dimensionalidad. Sin embargo, cuando los

modeladores consideran varios precios de acciones, índices o fondos cotizados en bolsa, enfrentarán matrices de varianzas y covarianzas de alta dimensionalidad. Aunque este no es el caso, es posible recomendar literatura para realizar análisis bajo estas circunstancias [92, 93, 94].

Hay 47 muestras (rendimientos mensuales) en los cinco escenarios analizados. La metodología utilizada para estimar los rendimientos y la matriz de covarianza se basa en una ventana móvil de 20 muestras de los rendimientos de los índices bursátiles (desde $t-19$ hasta t) utilizados en este estudio; se consideran las últimas 20 muestras. Entonces, en un escenario, se obtienen 27 portafolios, ya que se incluye cada nuevo dato (rendimientos) para calcular la estimación de los rendimientos y la matriz de covarianza. Considerando dichas estimaciones, se procede a optimizar los portafolios. La ventaja de usar este método radica en capturar posibles cambios a lo largo del tiempo en las estadísticas de los índices, lo que permite hacer una estimación dinámica que ayuda a mejorar la optimización del modelo [95]. Finalmente, se componen los rendimientos de cada modelo (portafolio) en todos los subconjuntos para evaluar cada modelo y obtener sus rendimientos totales.

4.2.2. Resolución de los problemas de optimización

Los modelos 1, 2 y 3 son no lineales y tienen una única función objetivo. Cabe destacar que, aunque el modelo 3 tiene como objetivos la minimización de la varianza y la maximización de la entropía, se utiliza el enfoque propuesto por [96] para resolverlo como un problema de un solo objetivo.

Para el modelo 1 se realiza la optimización minimizando la varianza sujeta a once diferentes valores de retorno, desde r_{min} hasta r_{max}, donde r_{min} es el retorno mínimo posible y r_{max} es el retorno máximo posible dado el intervalo bajo análisis. Este intervalo de retorno se recorrió con un paso de 0,1 veces la longitud del intervalo de retorno, generando 11 soluciones óptimas, una para cada retorno restringido. En el modelo 2, la maximización de la entropía del portafolio está sujeta a valores de retorno y varianza. El intervalo de retorno y varianza se recorrió con un paso de 0,1 veces la longitud del intervalo (de retorno y varianza), generando 121 soluciones (11 retornos x 11 varianzas).

En el caso del modelo 3, se minimizó la varianza y se maximizó la entropía en la función objetivo, donde se restringió el retorno; se implementó esta restricción como una igualdad y una desigualdad. El recorrido del intervalo de retorno se hizo con un paso de 0,1 veces la longitud del intervalo, generando 22 soluciones.

Los modelos 4 a 7 son no lineales y tienen múltiples funciones objetivo. Se utilizó la metodología de restricción-ϵ [97] para encontrar soluciones. Esta metodología resuelve un problema multiobjetivo minimizando una sola función mientras las demás se dejan

como restricciones. El enfoque de restricción-ϵ convierte las múltiples funciones objetivo en restricciones controladas por un parámetro en el lado derecho, ϵ. De esta manera, es posible estimar la frontera eficiente de soluciones realizando un análisis de sensibilidad en los parámetros ϵ. El concepto de frontera eficiente, en este caso, se refiere al problema de optimización con múltiples funciones objetivo, donde se busca encontrar el conjunto de soluciones óptimas que maximicen una función objetivo mientras se satisfacen las restricciones impuestas por las otras funciones objetivo.

En términos matemáticos, se considera un problema de optimización multiobjetivo general definido en la Ecuación (4.1), con $\{f_i(\mathbf{x})\}_{i=0}^{n}$ como funciones objetivo y $j \in \{1, ..., m\}$ restricciones, definidas por las funciones $g_j(\mathbf{x})$ y sus parámetros en el lado derecho, b_j. El enfoque de restricción-ϵ se muestra en la Ecuación (4.2), donde se mantiene $f_0(\mathbf{x})$ como una única función objetivo, y se considera $\{f_i(\mathbf{x})\}_{i=1}^{n}$ como restricciones con parámetros en el lado derecho $\{\epsilon_i\}_{i=1}^{n}$.

$$\begin{array}{lll} \underset{\{\mathbf{x}\geq 0\}}{\text{mín}} & \{f_0(\mathbf{x}), ..., f_n(\mathbf{x})\} & \\ \text{s.t.} & g_j(\mathbf{x}) \geq b_j & \forall j \in \{1, ..., m\} \end{array} \tag{4.1}$$

$$\begin{array}{lll} \underset{\{\mathbf{x}\geq 0\}}{\text{mín}} & f_0(\mathbf{x}) & \\ \text{s.t.} & g_j(\mathbf{x}) \geq b_j & \forall j \in \{1, ..., m\} \\ & f_i(\mathbf{x}) \geq \epsilon_i & \forall i \in \{1, ..., n\} \end{array} \tag{4.2}$$

Luego, se aplica el enfoque de restricción-ϵ para cada modelo indicado:

- **Modelo 4 - Máxima Entropía, Difuso en retorno y varianza**: Este modelo tiene dos funciones objetivo: la maximización de la entropía y el parámetro de control λ para la función de membresía, para la varianza y el retorno esperado del portafolio. Se mantiene la entropía como la única función objetivo en este caso, convirtiendo a λ en una restricción-ϵ. Se realizan los experimentos numéricos haciendo un análisis de sensibilidad en los ϵ correspondientes que delimitan λ. Estos intervalos están controlados por λ, que toma valores entre 0 y 1, recorriendo con un paso de 0,1, lo que genera 11 soluciones.

- **Modelo 5 - Maximización de la Entropía, entropía difusa en retorno, difuso en varianza**: Este modelo tiene tres funciones objetivo: la maximización de la entropía, la función entrópica difusa para el retorno esperado y el parámetro de control λ de la función de membresía para la varianza del portafolio. En este caso, se mantiene la entropía como la única función objetivo, convirtiendo las otras funciones objetivo en dos restricciones-ϵ. Se realizan los experimentos numéricos

haciendo análisis de sensibilidad en los parámetros ϵ correspondientes que delimitan λ y $g_1(\mathbf{x})$.

Utilizando el mismo enfoque del modelo 4, α y λ toman valores entre 0 y 1; α actúa sobre la restricción del retorno y λ modula el intervalo de varianza (trapezoidal), siendo dichos intervalos en realidad restricciones en el proceso de optimización. El recorrido de α y λ tiene un paso de 0,1, lo que genera 121 soluciones.

- **Modelo 6 - Maximización de la Entropía, entropía difusa en retorno y varianza**: Este modelo tiene tres funciones objetivo: la maximización de la entropía, la función entrópica difusa para el retorno esperado y las funciones de minimización de la varianza del portafolio. En este caso, se mantiene la entropía como la única función objetivo, convirtiendo las otras funciones objetivo en dos restricciones-ϵ. Se realizan los experimentos numéricos haciendo un análisis de sensibilidad en los parámetros ϵ correspondientes que delimitan respectivamente $g_1(\mathbf{x})$ y $\mathbb{V}(\mathbf{x})$. Siguiendo la misma metodología utilizada en los otros modelos, α es una variable que toma valores entre 0 y 1 y recorre este intervalo con un paso de 0,1, generando 11 soluciones.

- **Modelo 7 - Maximización de la Entropía y entropía difusa en retorno y varianza**: Este modelo tiene tres funciones objetivo: la maximización de la entropía y las funciones entrópicas difusas para el retorno esperado, y la varianza del portafolio. En este caso, se mantiene la entropía como la única función objetivo, convirtiendo las otras funciones objetivo en dos restricciones-ϵ. Se realizan los experimentos numéricos haciendo análisis de sensibilidad en los parámetros ϵ correspondientes que delimitan respectivamente $g_1(\mathbf{x})$ y $g_2(\mathbf{x})$. De esta manera, recorriendo estos parámetros con un paso de 0,1, se generan 121 portafolios.

El proceso de optimización de cada modelo termina cuando se selecciona el portafolio óptimo para cada uno (modelos 1 a 7) y calculamos sus rendimientos. Finalmente, se utiliza TOPSIS para seleccionar el mejor modelo en cuanto a sus tres características principales: retorno esperado, varianza y entropía. Además, se aplica un análisis de sensibilidad sobre la relevancia de cada variable para representar diferentes escenarios; por supuesto, la clasificación de los modelos cambiará en cada caso.

4.2.3. TOPSIS en detalle

La metodología TOPSIS es una metodología de decisión multicriterio que permite elegir la solución a un problema basado en múltiples atributos. Normalmente, los atributos son características opuestas, como el retorno y la volatilidad. La metodología determina la puntuación en función de la distancia euclidiana a un elemento ideal positivo y un elemento ideal negativo. Cuanto más cerca esté el elemento del elemento ideal positivo y más lejos esté del elemento ideal negativo, mejor será la puntuación. Estos elementos ideales se construyen tomando los valores más altos y más bajos de los elementos disponibles.

En este caso, se consideran los atributos para $k \in \{1, ..., K\}$ portafolios, incluidos sus retornos esperados, varianzas y entropías. El elemento ideal positivo es el que tiene el mayor retorno, la menor varianza y la mayor entropía. Por otro lado, el elemento ideal negativo es el que tiene el menor retorno, la mayor volatilidad y la menor entropía. Si se consideran los atributos $\mathbf{a}_k$ como se indica en la Ecuación (4.3), donde se calculan los valores de cada atributo considerando el portafolio $\mathbf{x}_k$ y su rendimiento basado en datos fuera de la ventana temporal.

$$\mathbf{a}_k = \left(r_p\left(\mathbf{x}_k\right) \ , \ \mathbb{V}\left(\mathbf{x}_k\right) \ , \ H\left(\mathbf{x}_k\right)\right) \tag{4.3}$$

Se realizan los siguientes pasos:

1. **Definir la matriz de rendimiento**: La matriz de rendimiento corresponde a la colección de atributos de los portafolios, constando de tres columnas (atributos), retorno, volatilidad y entropía.

2. **Normalización**: Este paso corresponde a convertir los valores de la matriz de rendimiento $\forall\ k \in \ \{1, .., K\}$ en valores entre 0 y 1:

$$\overline{\mathbf{a}_k} \ = \ \frac{\mathbf{a}_k}{\|\mathbf{a}_k\|} \ \ , \ \ \|\mathbf{a}_k\| \ = \ \sqrt{\sum_{l=1}^{3} a_{k,l}^2} \tag{4.4}$$

3. **Matriz normalizada ponderada**: Se multiplican las columnas de la matriz de rendimiento normalizada por un valor escalar, donde la suma de estos valores escalares es igual a 1.

$$\sum_{l=1}^{3} w_l \ = \ 1, \tag{4.5}$$

En la ecuación previamente expuesta, w_l son los escalares que multiplican las columnas; estos valores representan la relevancia relativa de los atributos. Consideramos múltiples configuraciones.

4. **Ideales positivo y negativo**: Se define el portafolio ideal positivo y negativo; usando los valores más altos y más bajos de los atributos de cada portafolio.

$$P = \left[\max_{k\in\{1,...,K\}} r_p(\mathbf{x}_k) \; ; \; \min_{k\in\{1,...,K\}} \mathbb{V}(\mathbf{x}_k) \; ; \; \max_{k\in\{1,...,K\}} H(\mathbf{x}_k) \right] \tag{4.6}$$

$$N = \left[\min_{k\in\{1,...,K\}} r_p(\mathbf{x}_k) \; ; \; \max_{k\in\{1,...,K\}} \mathbb{V}(\mathbf{x}_k) \; ; \; \min_{k\in\{1,...,K\}} H(\mathbf{x}_k) \right] \tag{4.7}$$

5. **Cálculo de la distancia**: Se calcula la distancia euclidiana a los portafolios ideales positivo y negativo ($\forall k \in \{1, ..., K\}$).

$$\begin{aligned} d_k^P &= \sqrt{\sum_{l=1}^{3}(a_{k,l} - p_l)^2} \\ d_k^N &= \sqrt{\sum_{l=1}^{3}(a_{k,l} - n_l)^2} \end{aligned} \tag{4.8}$$

Donde p_l y n_l son los componentes (retorno, varianza, entropía) de los elementos ideal positivo e ideal negativo, respectivamente.

6. **Cercanía relativa**: Se calcula la cercanía relativa a la solución ideal, que corresponde a la razón de la distancia al portafolio ideal negativo dividida por la suma de las distancias al portafolio ideal negativo y positivo.

$$C_k = \frac{d_k^N}{d_k^N + d_k^P} \tag{4.9}$$

7. **Resultados de la clasificación**: Se clasifica la solución desde la mayor cercanía relativa hasta la menor, según lo calculado en el paso anterior.

Según este enfoque propuesto, se consideran nueve configuraciones diferentes para los pesos con el fin de analizar las clasificaciones y discriminar qué modelos podrían ser mejores; estas configuraciones consideran el retorno como el más relevante. La principal razón por la que se tuvo que hacer esto es para representar que el retorno domina sobre

otras métricas a los ojos de los inversores cuando seleccionan su portafolio. Además, en la Tabla 4.8, se muestran los nueve conjuntos de pesos e_i para los atributos de los portafolios.

Cuadro 4.8: Configuraciones de pesos TOPSIS analizadas.

Peso de atributo / Escenario	1	2	3	4	5	6	7	8	9
Retorno [%]	33	50	50	50	70	60	60	70	60
Varianza [%]	33	50	0	25	15	40	0	30	30
Entropía [%]	33	0	50	25	15	0	40	0	10

4.3. Modelo de aprendizaje reforzado multiagente

El modelo de aprendizaje reforzado en esta investigación es una extensión significativa de los enfoques tradicionales de optimización de portafolios al integrar tanto la maximización de la rentabilidad y su ajuste por riesgo, como la minimización de la correlación entre diversos portafolios en una función de recompensa conjunta. Los agentes aprenden a tomar decisiones de inversión que no solo consideran la información del mercado, sino también las acciones de otros agentes, creando un entorno más realista para la optimización de portafolios en mercados complejos y dinámicos. Cabe señalar que al incluir la minimización de la correlación entre los portafolios y modelar estrategias frente a las perturbaciones externas, se permite que los agentes desarrollen estrategias de inversión robustas y diversificadas, que maximizan su rendimiento a nivel individual y global.

El modelo multiagente que se presenta es fundamentalmente diferente de los enfoques tradicionales, ya que los agentes no solo basan sus decisiones en la información del mercado, sino también en las estrategias que otros agentes desarrollan y adaptan. Este enfoque se aleja de los modelos puramente competitivos o cooperativos, y refleja de manera más precisa el comportamiento en los mercados financieros reales, donde los agentes no están completamente alineados, pero tampoco buscan activamente sabotear las estrategias de los demás.

Cada agente actúa de forma independiente para maximizar su propia utilidad y minimizar su riesgo, pero al mismo tiempo, los portafolios que desarrollan están sujetos a las acciones de otros agentes y a la interacción del entorno, del mercado en su conjunto. Las decisiones de inversión de un agente influyen en las decisiones de otros, lo que genera un entorno dinámico en el que los agentes deben aprender a adaptarse y desarrollar estrategias que puedan resistir las perturbaciones causadas por otros participantes del mercado.

Este entorno, donde los agentes interactúan pero no compiten directamente, hace

alusión a las herramientas de teoría de juegos, particularmente al concepto de equilibrio de Nash, en el cual los agentes alcanzan un estado donde ninguno tiene incentivos para desviarse unilateralmente de su estrategia. Este equilibrio es especialmente importante en los mercados financieros, donde la capacidad de anticipar y adaptarse a las acciones de otros inversores puede marcar la diferencia entre una estrategia de inversión exitosa o una fallida [98].

Uno de los retos más importantes que se aborda en este modelo es la minimización de la correlación entre los portafolios de los diferentes agentes. En un mercado real, los inversores intentan diversificar sus activos para reducir el riesgo sistemático (riesgo que afecta a todo el mercado) y minimizar la correlación entre los activos que poseen. En el modelo multiagente, la función de recompensa no solo maximiza el ratio de Sharpe individual de cada agente, sino que también incluye un término que penaliza las altas correlaciones entre los portafolios de los diferentes agentes. Este enfoque garantiza que los agentes busquen activamente estrategias de inversión diferenciadas, minimizando así la dependencia mutua entre portafolios y promoviendo una diversificación efectiva del riesgo.

El desafío consiste en garantizar que las estrategias aprendidas por los agentes sean robustas y estables no solo en condiciones normales de mercado, sino también ante perturbaciones externas o cambios inesperados en el comportamiento de otros agentes. Esta punto es esencial en finanzas, donde los mercados pueden experimentar periodos de alta volatilidad, crisis financieras o cambios abruptos en las condiciones del mercado.

Este modelo entonces, además de garantizar la estabilidad y la convergencia de las políticas aprendidas, debe, a través de los agentes, alcanzar un equilibrio estable que no solo sea localmente óptimo, sino que también sea resistente a perturbaciones y cambios en las estrategias de otros agentes convergiendo a un optimo global, entendida como la cartera de la gestora.

4.3.1. Entorno

El entorno se formará por medio de la información estadística disponible y detallada en capítulos anteriores , en un escenario donde coexisten elementos competitivos e individualistas, costos de transacción y aparecen recompensas entendidas como las rentabilidades o retornos sujetos al riesgo individual de cada agente. Además de elementos cooperativos, como la penalización por correlación, lo que fomenta la diversificación entre los agentes, para maximizar la eficiencia del conjunto. De esta manera, los datos históricos incluyen los siguientes precios para cada acción i en el día t:

- Open_t^i: Precio de apertura.

- High_t^i: Precio máximo del día.
- Low_t^i: Precio mínimo del día.
- Close_t^i: Precio de cierre.

Estos precios permiten calcular los retornos diarios, que se definen como:

$$r_t^i = \frac{\text{Close}_t^i - \text{Close}_{t-1}^i}{\text{Close}_{t-1}^i} \tag{4.10}$$

Siendo r_t^i el retorno de la acción i en el día t. Para esto, se consideran ventanas móviles, técnica que permite al agente observar los precios de los últimos n días antes de tomar una decisión de trading. Además, se considera el tamaño de la ventana $n = 40$, entonces en el paso t, el agente observará los precios desde el día $t - 39$ hasta el día t.

El estado S_t, que recibe el agente en el paso t, está compuesto por una serie temporal de los últimos n días de precios normalizados para cada acción i. Para cada acción, el estado S_t^i se representa como una matriz de dimensiones $4 \times n$, donde las filas corresponden a los precios de apertura, máximo, mínimo y cierre, respectivamente:

$$S_t^i = \begin{pmatrix} \frac{\text{Open}_{t-n+1}^i}{\text{Close}_t^i} & \cdots & \frac{\text{Open}_t^i}{\text{Close}_t^i} \\ \frac{\text{High}_{t-n+1}^i}{\text{Close}_t^i} & \cdots & \frac{\text{High}_t^i}{\text{Close}_t^i} \\ \frac{\text{Low}_{t-n+1}^i}{\text{Close}_t^i} & \cdots & \frac{\text{Low}_t^i}{\text{Close}_t^i} \\ \frac{\text{Close}_{t-n+1}^i}{\text{Close}_t^i} & \cdots & 1 \end{pmatrix} \tag{4.11}$$

Cada fila representa una serie temporal de precios normalizados. Además, la columna final es siempre 1, ya que el precio de cierre actual se normaliza respecto a sí mismo. Conforme el tiempo avanza, la ventana móvil se desplaza un día hacia adelante, eliminando el día más antiguo y añadiendo el nuevo día al final. Esto asegura que el agente siempre esté tomando decisiones basadas en la información más reciente disponible.

Para asegurar que los datos de entrada estén en una escala comparable, los precios se normalizan en función del precio de cierre del día actual. Los precios se transforman de la siguiente manera:

$$\text{Open}_t^i = \frac{\text{Open}_t^i}{\text{Close}_t^i}, \quad \text{High}_t^i = \frac{\text{High}_t^i}{\text{Close}_t^i}, \quad \text{Low}_t^i = \frac{\text{Low}_t^i}{\text{Close}_t^i}, \quad \text{Close}_t^i = 1 \tag{4.12}$$

Esto permite que los precios de apertura, máximo y mínimo se expresen como propor-

ciones del precio de cierre, lo que facilita que los modelos trabajen con valores normalizados. La ventana móvil le proporciona al agente contexto histórico que le permite identificar patrones temporales en los precios, siendo algo esencial para que el agente pueda:

- **Detectar tendencias**: Si los precios han subido o bajado de manera consistente.
- **Identificar volatilidad**: Fluctuaciones grandes entre precios máximos y mínimos.
- **Reconocer patrones de precios**: Como rompimientos o reversión de precios.

4.3.2. Agentes

El sistema multiagente está diseñado para abordar los diferentes segmentos del mercado financiero: renta fija nacional, renta fija internacional, renta variable nacional y renta fija internacional. En este contexto, cada segmento será gestionado por un agente especializado. A continuación, se explica el enfoque para cada uno.

En primer lugar, se comprende el agente para la renta fija nacional. Este agente se encargará de modelar las inversiones en instrumentos de renta fija emitidos en Chile, como bonos del Banco Central, letras hipotecarias y bonos corporativos. El objetivo de este agente será optimizar el rendimiento ajustado por riesgo de los instrumentos de deuda, considerando el bajo riesgo que tradicionalmente caracteriza a este tipo de inversiones.

La inclusión de las medidas de duración y convexidad en un modelo para renta fija es fundamental porque ambas características capturan la sensibilidad de los precios de los bonos frente a variaciones de los tipos de interés. La duración estima de forma lineal la variación porcentual del precio ante un cambio marginal en la tasa, proporcionando al agente de RL una señal clara de exposición al riesgo de mercado. Sin embargo, la relación precio–tasa no es estrictamente lineal: la convexidad mide la curvatura de esa relación, permitiendo corregir las aproximaciones de la duración y gestionar correctamente las asimetrías. Al incorporar tanto la diferencia de duración como la convexidad en la función de recompensa, penalizando excesiva sensibilidad lineal o no lineal, el agente aprende políticas de asignación que equilibran rendimiento y protección frente a cambios bruscos de tasas. De este modo, el modelo refuerza decisiones de inversión más estables en portafolios de renta fija, mejorando la gestión del riesgo de tasas y optimizando el perfil de retorno ajustado.

El agente usará un enfoque conservador, buscando estabilidad en los rendimientos. Entonces, para la construcción de este modulo de aprendizaje reforzado profundo uniagente y multiagente se empleará un algoritmo en DDPG siendo el agente neuronal de tipo MLP (Multi-Layer Perceptron), donde el objetivo será maximizar el retorno, ajustando también

por riesgo y minimizar costos de movimientos en el portafolio. Formado por instrumentos que generalmente son estables pero pueden reaccionar a cambios macroeconómicos, como las políticas del Banco Central de Chile. La estabilidad de la renta fija y la predictibilidad de sus pagos hacen que este agente se enfoque más en la reducción de la volatilidad y la duración (sensibilidad a tasas de interés), que en la rentabilidad pura.

En segundo lugar, se comprende el agente para renta fija internacional. Este agente será responsable de manejar las inversiones en bonos emitidos por gobiernos y corporaciones extranjeras. Es importante notar que dichos instrumentos suelen tener una menor volatilidad que las acciones internacionales, pero presentan el desafío de exposición al tipo de cambio, a riesgos políticos y económicos en otras regiones.

Entonces, acorde al capítulo 3 el enfoque para este agente es el uso de modelos Actor-Crítico, específicamente se utilizará PPO, adecuado para optimizar en entornos continuos y de alta volatilidad, ya que mantiene la estabilidad en las actualizaciones de la política. Esto es ideal para gestionar el riesgo en mercados internacionales, permitiendo evaluaciones continuas de riesgo-retorno en función de cambios económicos globales donde el Actor selecciona las acciones (en este caso, los bonos o instrumentos de renta fija), y el Crítico evalúa la calidad de esas decisiones basándose en la relación entre el retorno y el riesgo. Este agente neuronal será del tipo MLP y buscará maximizar el rendimiento ajustado al riesgo mientras minimiza los costos asociados a las fluctuaciones del tipo de cambio.

En tecer lugar, el agente para la renta variable nacional debe estar especializado en el mercado de acciones chilenas, operando con los datos de precios de apertura, cierre, máximo y mínimo de las empresas que cotizan en la Bolsa de Comercio de Santiago. Su objetivo será identificar oportunidades de crecimiento en el mercado local, a través del análisis de tendencias y patrones de los precios históricos.

Este agente considerado como una red LSTM (Long Short-Term Memory) que permite capturar las dependencias temporales en los datos históricos. buscando maximizar el crecimiento del capital a largo plazo, invirtiendo en acciones con alto potencial de apreciación en el mercado nacional.

Además, el uso de LSTM permite al agente gestionar mejor los datos históricos y realizar predicciones de rendimiento y volatilidad con base en patrones observados en el mercado. En conjunto con un algoritmo DDPG no solo maximiza el sharpe, sino además identifica activos con potencial de crecimiento sostenido, gestionando el riesgo inherente a las acciones nacionales y de paso deberá minimizar costos de transacción.

En cuarto lugar, el agente para la renta variable internacional gestionará las inversiones en acciones de empresas que cotizan en bolsas internacionales, como la Bolsa de Nueva

York (NYSE) o la Bolsa de Londres (LSE). Para esto, utilizará un enfoque que permita diversificar la exposición geográfica y sectorial del portafolio.

Dado que la renta variable internacional implica mayor volatilidad, este agente se modelará con una red CNN (Convolutional Neural Network) para extraer patrones significativos de los datos de precios de múltiples mercados internacionales. En este contexto, el agente estará diseñado para encontrar oportunidades de inversión en acciones globales que ofrezcan altos rendimientos, pero que también conlleven un mayor riesgo. Además, este agente debe gestionar un portafolio global diversificado, aprovechando oportunidades de alto crecimiento en sectores clave, mientras minimiza los riesgos asociados a eventos geopolíticos o fluctuaciones de divisas.

En este contexto el sistema se basa en modelos uniagente y en otro modelo multiagente que incopora los primeros pero bajo supervisión. Cada agente tendrá su propio objetivo individual, pero al mismo tiempo, estarán colaborando para lograr una optimización global del portafolio. Los agentes trabajarán en conjunto para optimizar la asignación de capital en el portafolio. Aunque cada agente está especializado en un segmento del mercado, sus decisiones afectarán el desempeño global, para evitar la sobreexposición a ciertos riesgos, los agentes deberán coordinarse manteniendo un adecuado equilibrio entre los diferentes tipos de activos.

En términos de coordinación y evaluación global, se detalla nuevamente que el sistema multiagente tiene una capa adicional de supervisión que evalúa el rendimiento conjunto de los agentes y ajusta las ponderaciones del portafolio RL de cada agente de acuerdo con los objetivos generales del inversor. Esta capa actua como una “meta-red”, coordinando a los agentes de renta fija y variable, tanto a nivel nacional como internacional, para garantizar que las decisiones de cada uno estén alineadas con el rendimiento esperado global.Es importante notar que la meta-red no es un agente independiente. Esto hace referencia a que no opera como un agente más dentro del sistema, ni toma decisiones de inversión en un segmento particular, sino que ajusta y coordina las asignaciones de capital para optimizar el rendimiento global.

En este contexto, actúa mediante una Función de Supervisión: supervisa las actividades de cada agente y ajusta las ponderaciones del portafolio en función de los resultados históricos y las proyecciones futuras. Es decir, mientras los agentes se centran en optimizar su segmento de mercado, la meta-red asegura que las decisiones de cada uno mantengan el equilibrio del portafolio global.

La meta-red se implementa a través de técnicas de aprendizaje por refuerzo profundo para ajustar dinámicamente las asignaciones entre los agentes. En base a esto, actor-crítico se emplea para evaluar las decisiones de cada agente en función de su impacto en la meta

general del portafolio. Estos ajustes buscan minimizar la exposición a riesgos específicos y maximizar la eficiencia del portafolio completo, lo que permite que cada agente trabaje en colaboración hacia el objetivo global, en lugar de operar de manera aislada.

A continuación, se muestra la Tabla 4.9 que contiene los agentes y algoritmos dentro del sistema multiagente para la gestión de portafolios:

Cuadro 4.9: Agentes y algoritmos en el sistema multiagente.

Segmento	**Agente**	**Red Neuronal**	**Algoritmo**
Renta Fija Nacional	Agente de Renta Fija Nacional	MLP	DDPG
Renta Fija Internacional	Agente de Renta Fija Internacional	MLP	PPO
Renta Variable Nacional	Agente de Renta Variable Nacional	LSTM	DDPG
Renta Variable Internacional	Agente de Renta Variable Internacional	CNN	DDPG
Supervisión Global	Meta-Red	MLP	Actor-Crítico Modificado (DDPG)

4.3.3. Función de recompensa

En el caso de RFN la función de recompensa R_t en el tiempo t maximiza el retorno, el rendimiento ajustado por riesgo y controla la sensibilidad del portafolio ante fluctuaciones en las tasas de interés a través de la convexidad y la duración , además de los costos de transacción del portafolio del área. Los términos clave de esta función son:

$$(FR)_t = \lambda_1 \cdot R_t + \lambda_2 \cdot \frac{E[R_t] - R_f}{\sigma_i(t)} - \lambda_3 \cdot |D_t - D_{t-1}| + \lambda_4 \cdot |C_t - C_{t-1}| - \lambda_5 \cdot \tau \cdot \sum_{k=1}^{n} |w_{k,t} - w_{k,t-1}| \tag{4.13}$$

Donde:

- **Retorno** $(\lambda_1 \cdot E[R_t])$: Maximiza el retorno del portafolio a nivel diario.
- **Sharpe Ratio** $\left(\lambda_2 \cdot \frac{E[R_t]-R_f}{\sigma_i(t)}\right)$: Maximiza el retorno ajustado por riesgo $\left(\sigma_{\mathrm{i}}(t) = \sqrt{\sum_k \sum_l w_k w_l \cdot \mathrm{Cov}(R_k, R_l)}\right)$ del agente i.
- **Penalización por cambios de Duración** $(\lambda_3 \cdot |D_{\mathrm{t}} - D_{t-1}|)$: Minimiza la desviación de la duración del portafolio . La duración mide la sensibilidad del portafolio

ante cambios en las tasas de interés. Para limitar esta sensibilidad, se aplica una penalización cuando la duración del portafolio, D_{t-1}, cambia de un paso de tiempo a otro D_{t}.

- **Penalización por cambios de Convexidad** ($\lambda_4 \cdot |C_{\mathrm{t}} - C_{t-1}|$): La convexidad refleja la estabilidad del portafolio frente a cambios bruscos en las tasas de interés. Para mantener esta estabilidad, se agrega una penalización basada en la desviación de la convexidad del portafolio C_t respecto a un objetivo $C_{\text{t-1}}$.

  Cabe señalar que la función de recompensa podría ser modificada a fin de reforzar la convexidad, el objetivo detrás es que la política aprende a privilegiar instrumentos con mayores curvaturas en la relación precio–tasa, mejorando la protección ante movimientos extremos sin sacrificar excesivo rendimiento. En ese caso la función ya no conlleva una penalización si no un premio en búsqueda de la maximización de la función ($\lambda_4 \cdot |C_{\mathrm{t}}|$)

- **Penalización por Costos de Transacción** ($\lambda_5 \cdot \tau \cdot \sum_{k=1}^{n} |w_{k,t} - w_{k,t-1}|$): Minimiza el impacto de cambios frecuentes en la composición del portafolio.

Donde:

- $\mathrm{Cov}(R_t^i, R_t^j)$: Covarianza entre los retornos de los activos i y j.
- n: Número total de activos.
- σ_{p}: Desviación estándar o volatilidad del portafolio para el agente.
- $w_{k,t}$: Peso del activo k en el portafolio en el tiempo t.
- $w_{k,t-1}$: Peso del activo k en el portafolio en el tiempo $t-1$.
- λ_1: Peso que controla la importancia del retorno del portafolio en la recompensa total.
- λ_2: Peso que ajusta la importancia del rendimiento o índice de Sharpe del portafolio.
- λ_3: Peso que ajusta la influencia de la estabilidad y la duración.
- λ_4: Peso que ajusta la importancia de la convexidad.
- λ_5: Peso asignado a la penalización de costos de transacción.
- τ: Costo asociado con el cambio en la ponderación de los activos

En el caso de cada agente RFI, RVN, RVI la función de recompensa se compone de

tres factores, la maximización del retorno del portafolio del agente, la máximizacion del índice de Sharpe del portafolio de cada agente y la penalización por costos de transacción.

- **Retorno** ($\lambda_1 \cdot E[R_t^i]$): Maximiza el retorno del portafolio diario del agente i.
- **Sharpe Ratio** $\left(\lambda_2 \cdot \frac{E[R_t^i]-R_f}{\sigma_i(t)}\right)$: Maximiza el retorno ajustado por riesgo $\left(\sigma_{\mathrm{i}}(t) = \sqrt{\sum_k \sum_l w_k w_l \cdot \mathrm{Cov}(R_k, R_l)}\right)$ del agente i.
- **Penalización por Costos de Transacción** $\left(\lambda_5 \cdot \tau \cdot \sum_{k=1}^{n} \left|w_{k,t}^i - w_{k,t-1}^i\right|\right)$: Minimiza el impacto de cambios frecuentes en la composición del portafolio del agente i.

La recompensa para cada agente i en el tiempo t queda definida por:

$$(FR)_t^i = \lambda_1 \cdot \mathrm{R}_t^i + \lambda_2 \cdot \frac{E[R_t^i] - R_f}{\sigma_i(t)} - \lambda_3 \cdot \tau \cdot \sum_{k=1}^{n} |\mathrm{w}_{k,t}^i - \mathrm{w}_{k,t-1}^i| \tag{4.14}$$

Donde:

- R_t^i: Retorno del portafolio del agente i en el tiempo t,
- R_f: Tasa libre de riesgo.
- $\sigma_{\mathrm{i}}(t)$: Volatilidad o desviación estándar de los retornos del portafolio del agente i en (t).
- $\mathrm{Cov}(R_t^i, R_t^j)$: Covarianza entre los retornos de los activos i y j.
- $\mathrm{w}_k^i(t)$: Proporción del activo k en el portafolio del agente i en el tiempo t,
- τ: Costo asociado con el cambio en la ponderación de los activos.
- $\lambda_1, \lambda_2, \lambda_3$: Coeficientes de ponderación que determinan la importancia de cada componente en la función de recompensa.

Para finalizar en el contexto multiagente, la función de la meta-red puede representarse mediante una función de recompensa global que evalúa el rendimiento conjunto de los agentes. Esto mediante en tres factores claves, siendo representada en forma vectorial, donde los pesos y los factores se expresan como vectores.

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} \text{Return}_{\text{global}} \\ \text{Sharpe Ratio}_{\text{global}} \\ \text{Agents Corr.} \\ TC_{\text{global}} \end{bmatrix} \tag{4.15}$$

Entonces, la función de recompensa global de la meta-red R_{meta} se expresa como:

$$R_{\text{meta}} = \boldsymbol{\lambda} \cdot \mathbf{X} = \lambda_1 \cdot \text{Return}_{\text{global}} + \lambda_2 \cdot \text{Sharpe Ratio}_{\text{global}} - \lambda_3 \cdot |\text{Agents Corr.}| - \lambda_4 \cdot TC_{\text{global}} \tag{4.16}$$

En términos de la maximización del índice de Sharpe global, es la medida del rendimiento ajustado al riesgo del portafolio global, donde los activos son los retornos de los portafolios por área y se define como:

$$\text{Sharpe Ratio}_{\text{global}} = \frac{E[R_{\text{global}}] - R_f}{\sigma_{\text{global}}} \tag{4.17}$$

Donde:

- $E[R_{\text{global}}]$: Rendimiento esperado global, calculado como:

$$E[R_{\text{global}}] = \sum_i w_i \cdot E[R_i] \tag{4.18}$$

- R_f: Tasa libre de riesgo.
- σ_{global}: Desviación estándar global del portafolio, calculada como:

$$\sigma_{\text{global}} = \sqrt{\sum_i \sum_j w_i w_j \cdot \text{Cov}(R_i, R_j)} \tag{4.19}$$

En términos de la minimización, se debe minimizar el módulo de la correlación entre portafolios de los agentes i y j a través de los retornos de cada área y se calcula utilizando la correlación de Pearson:

$$\text{Corr}(i, j) = \frac{\text{Cov}(R_t^i, R_t^j)}{\sigma(R_t^i) \cdot \sigma(R_t^j)} \tag{4.20}$$

Donde:

- $\text{Corr}(i, j)$: Correlación entre los retornos de los portafolios de los agentes i y j.
- $\text{Cov}(R_t^i, R_t^j)$: Covarianza entre los retornos de los portafolios de los agentes i y j,
- $\sigma(R_t^i)$ y $\sigma(R_t^j)$: Desviaciones estándar de los retornos de los portafolios.

Los términos clave de esta función de recompensa (cabe notar n=4, bajo las mismas variables predefinidas, quedan descritos a continuación:

$$R_{\text{meta}} = \lambda_1 \cdot R_t + \lambda_2 \cdot \frac{E[R_g] - R_f}{\sigma_g} + \lambda_3 \cdot \frac{1}{2 \cdot (n-1) \cdot n} \sum_{i=1, i>j}^{n} \sum_{j=1}^{n} |\text{Corr}(i, j)| - \lambda_4 \cdot \frac{\tau}{n-1} \cdot \sum_{i=1}^{n} \sum_{k=1}^{n_i} |w_{k,t} - w_{k,t-1}| \tag{4.21}$$

Donde:

- n: Número total de agentes.(4)
- σ_{g}: descviación estándar portafolio global.
- $w_{k,t}$: Peso del activo (Agente) k en el portafolio en el tiempo t.
- $w_{k,t-1}$: Peso del activo k en el portafolio en el tiempo $t-1$.
- λ_1: Peso que controla la importancia del Retorno en la recompensa total.
- λ_2: Peso que controla la importancia del Sharpe Ratio en la recompensa total.
- λ_3: Peso asignado a la penalización de la correlación entre los retornos de los portafolios de cada agente (i agentes).
- λ_4: Peso asignado a la penalización de costos de transacción.

4.3.4. Consideraciones de estabilidad y convergencia en el modelo MARL

En este esquema, cada agente mantiene autonomía operativa pero recibe periódicamente señales de la política global mediante soft updates de sus redes objetivo.

- **Actualización suave de parámetros:**

$$\theta^{i,\text{target}} \leftarrow \tau\, \theta^{\text{supervisor}} + (1-\tau)\, \theta^{i,\text{target}}, \quad 0 < \tau \ll 1. \tag{4.22}$$

A partir de la ecuación anterior, se tiene que θ^i son los pesos de la red (actor o crítico) del agente i, y $\theta^{\text{supervisor}}$ los de la red meta.

- **Frecuencia de sincronización:**
  - Cada K episodios locales, se realiza un episodio global de coordinación.
  - Durante ese episodio, se actualizan las τ-mezclas de pesos para integrar gradualmente la señal global.

Para garantizar convergencia y evitar comportamientos oscilatorios o colapsos en el entrenamiento, se considera:

1. **Buffers de experiencia independientes**: Cada agente y el supervisor mantienen su propio buffer con transiciones (s_t, a_t, r_t, s_{t+1}). Al separar los recuerdos de cada nivel, evitamos que muestras muy sesgadas o trayectorias típicas del supervisor que contaminen el aprendizaje local de los agentes. Esto preserva la coherencia de la distribución de estados y acciones de cada red y permite ajustar de forma independiente la frecuencia de actualización de cada buffer según las necesidades de convergencia de cada actor o crítico.

2. **Normalización de recompensas**: Los valores de recompensa local R_t^i y global R_t^{meta} pueden tener escalas muy diferentes, dadas por la suma de los retornos. Para evitar que una señal muy grande y que domine el gradiente, se re-escala cada R_t a un rango fijo $[-1, 1]$ usando estadísticas móviles de media y desviación estándar. Esta técnica estabiliza la magnitud de las actualizaciones y reduce la sensibilidad a valores extremos.

3. **Control de magnitud de τ y tasas de aprendizaje**:
   - **Soft-updates**: $\tau \in [0{,}001, 0{,}01]$ para mezclar gradualmente los pesos del supervisor en las redes objetivo de los agentes, evitando saltos bruscos en la política
   - **Learning rates diferenciadas**: $\alpha_{\text{critic}} = 10^{-3}$ y $\alpha_{\text{actor}} = 10^{-4}$, de modo que la estimación de valores se estabilice antes de actualizar la política.

4. **Monitoreo dual de convergencia**: Se debe evaluar períodicamente el índice de Sharpe local de cada agente y verificar mejora en rendimiento ajustado al riesgo, junto con el índice de Sharpe y correlación global del supervisor para asegurar diversificación y retorno ajustado. Si alguna métrica se estanca o empeora, se ajustan los hiperparámetros (learning rates, τ, tamaño de buffer, etc.) para restaurar un

ritmo de aprendizaje adecuado. Se detendrá el entrenamiento cuando el rendimiento conjunto deje de mejorar, preservando la capacidad de generalización.

5. **Resumen Hiperparámetros para los modelos**

   Para asegurar un aprendizaje eficiente y estable de cada agente, se define un conjunto específico de hiperparámetros y arquitecturas neuronales, de acuerdo con la naturaleza de cada clase de activo y el rol desempeñado por el agente. La siguiente tabla resume las elecciones algorítmicas, tasas de aprendizaje, tipos de red y detalles arquitectónicos empleados para cada componente del sistema; tanto los agentes individuales como el supervisor global. Estos parámetros fueron seleccionados en base a evidencia empírica en entornos financieros y buenas prácticas en aprendizaje reforzado profundo, considerando las prácticas comunes en la literatura y experimentación preliminar.

Cuadro 4.10: Hiperparámetros y arquitecturas de los componentes RL

Parámetro	**Agente RFN**	**Agente RFI**	**Agente RVN**	**Agente RVI**	**Supervisor Global**
Algoritmo	DDPG	PPO	DDPG	DDPG	Actor-Critic (PPO/DDPG)
Tasa de Aprendizaje (Actor)	1e-4	3e-4	1e-4	1e-4	1e-4
Tasa de Aprendizaje (Crítico)	1e-3	1e-3	1e-3	1e-3	1e-3
Factor de Descuento (γ)	0.99	0.99	0.99	0.99	0.99
Actualización Suave (τ)	0.005	N/A	0.005	0.005	0.005
Tamaño de Lote	128	2048	128	128	256
Tamaño del Búfer de Repetición	10^6	N/A	10^6	10^6	10^6
Tipo de Red	MLP	MLP	LSTM	CNN	MLP
Capas Ocultas	2	2	1 LSTM + 1 Densa	2 conv + 1 densa	2
Neuronas por Capa	256	64	256	128	256
Filtros/Kernels CNN	N/A	N/A	N/A	32 filtros, kernel=3	N/A
Funciones de Activación	ReLU	Tanh	ReLU	ReLU	ReLU

4.3.5. Representación esquemática

Cada agente se conecta con el entorno local representado por un cuadro vacío. Este cuadro representa el espacio donde el estado compartido de la meta-red global, denotado por S_t, se convierte en un estado específico para cada agente, simbolizado como s_t. Esta representación abstracta, componente fundamental del modelo permite visualizar cómo cada agente accede a la información general de la meta-red a través de un entorno propio, antes de tomar decisiones, es decir,permite que los agentes conviertan el estado global en información específica y relevante para su propio contexto de optimización.

El estado compartido S_t incluye información relevante para todos los agentes: el rendimiento global, la correlación entre portafolios y los costos de transacción globales o conjuntos. Esta información es entregada a cada agente mediante el entorno local, donde se filtra y transforma en un estado específico s_t, adaptado a las características y necesidades del área particular de cada agente, como la renta fija nacional o la renta

variable internacional.

El proceso de conversión de S_t a s_t es clave en el funcionamiento del modelo multiagente:

1. **Recepción de** S_t: Cada agente accede a la información global proporcionada por la meta-red en forma de S_t.

2. **Conversión a** s_t: El estado compartido S_t se adapta proporcionalmente su peso a nivel de activo y no de agente en el entorno global, para convertirse en un estado específico s_t en el entorno local, que refleja las características únicas de cada agente. Por ejemplo, un agente enfocado en renta fija ajustará S_t para centrarse en la estabilidad de bonos, mientras que un agente de renta variable interpretará S_t para evaluar tendencias y volatilidad en acciones.

3. **Toma de decisiones**: Con base en s_t, cada agente decide una acción A_t optimizada para su área particular, logrando sus objetivos individuales de maximización de rendimiento y minimización de riesgo, en coordinación con el objetivo global de la meta-red.

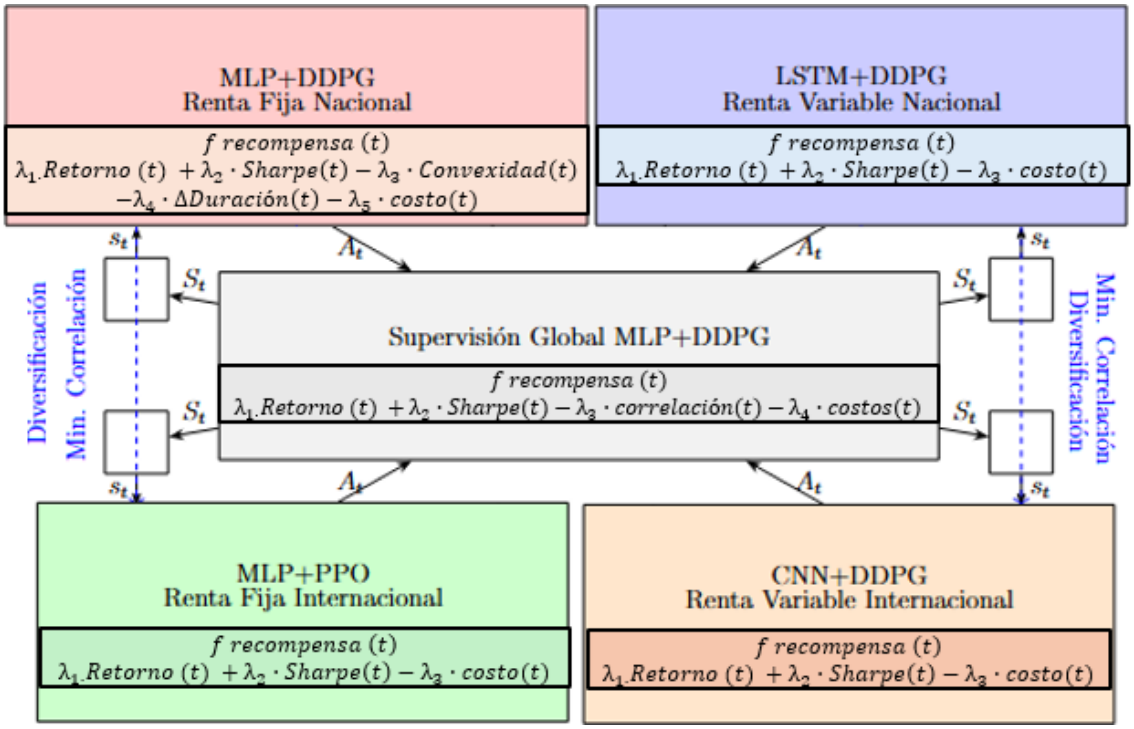

Figura 4.7: Esquema multiagente para la gestión y optimización de portafolios.

Capítulo 5

Resultados y análisis

5.1. Resultados preliminares

La presente sección tiene por objetivo evidenciar mediante modelos clásicos los supuestos básicos de esta tesis, se analiza la frontera eficiente construida a través de la teoría de Markowitz para una cartera representativa de la bolsa americana a través de algunas de las acciones más transadas y, por otro lado, de los índices más representativos del PIB mundial. Además, se encuentra el conjunto de portafolios conformados por todas las combinaciones de riesgo–rendimiento (Sharpe) que se pueden obtener entre los distintos activos que hacen parte de dicho portafolio y que ofrecen el rendimiento esperado más alto para cualquier nivel de riesgo dado y viceversa. Es importante notar que no existen limitaciones para la creación de portafolios, ya que estos se ajustan a los criterios de rentabilidad y riesgo de cada inversionista.

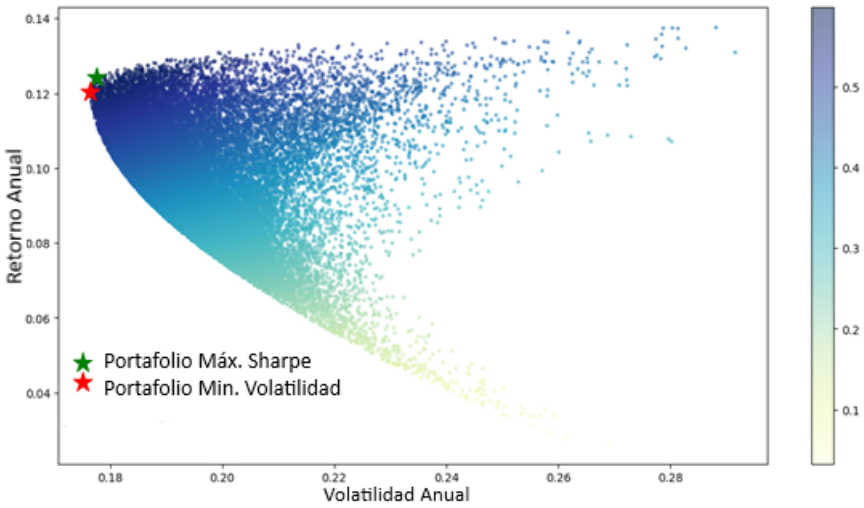

Figura 5.1: Frontera Eficiente para índices.

En la Figura 5.1 se observa que los portafolio de mínima varianza y máximo sharpe ratio están muy cerca. Esto fenómeno ocurre debido a que en el periodo seleccionado el índice accionario (S&P 500 de US) es el de mejor retorno y menor volatilidad, lo que va en contra de la teoría económica que sostiene que a mayor retorno mayor volatilidad.

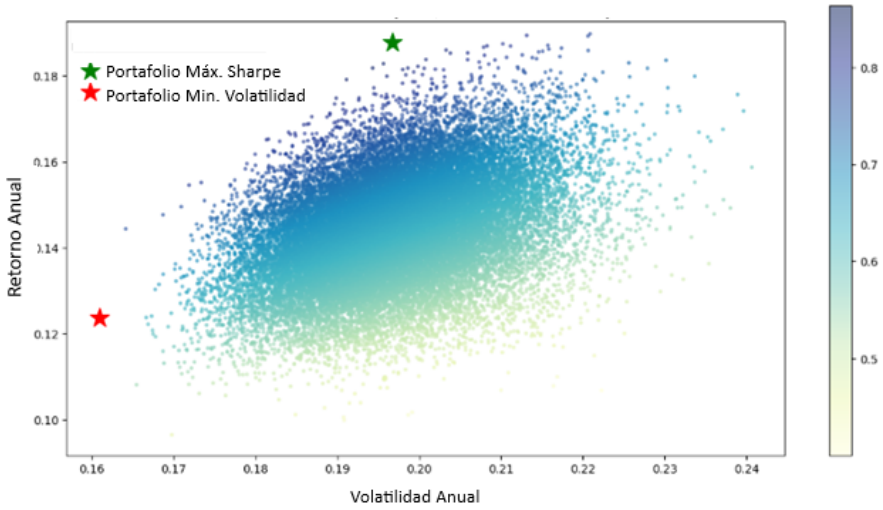

Figura 5.2: Frontera Eficiente para acciones de la bolsa americana.

Por otro lado, en el gráfico de la Figura 5.2 se muestran acciones de la bolsa americana. En este caso, la nube de puntos (portafolios) sigue la lógica de a mayor retorno mayor volatilidad. Se observa que, tanto el portafolio de mínima volatilidad y máximo sharpe ratio, están en la frontera eficiente del universo de portafolios. El fenómeno captado por los índices es uno de los fenómenos llamativos y objetivo de solución de esta tesis, debido a que es donde los modelos tradicionalmente utilizados fallan, y por ende, donde se pretende evidenciar el potencial uso del aprendizaje reforzado como técnica de optimización de portafolios.

El artículo de [99],explica algunos conceptos teóricos básicos para el inversor minorista, como el valor nominal, el valor de mercado, el valor de rescate, el rendimiento y el rendimiento al vencimiento de un bono. Se centra en la relación inversa entre el tipo de interés y el precio del bono, e ilustra la hipótesis de las expectativas, la teoría del mercado segmentado y la teoría del hábitat preferente. El bono se considera un instrumento de ahorro fiscal.

En[100, 101, 102], se especifica que las estrategias basadas en índices, implementadas a través de ETFs y futuros de índice, se han consolidado como el mecanismo predominante para gestionar carteras debido a su liquidez profunda, diversificación automática y capacidad de escalar grandes volúmenes con bajos costos de transacción. Estos vehículos pasivos reducen el riesgo asociado a la selección de acciones individuales y facilitan la replicación

de amplios universos de activos.

Sin embargo, de acuerdo a [103, 104] la investigación académica convencional ha seguido centrada en modelos de valoración de activos y en la prueba de factores de riesgo bajo supuestos ideales de mercado, sin abordar de manera sistemática la ejecución real de órdenes sobre índices ni el análisis empírico de deslizamiento y provisión de liquidez. Tradicionalmente consideran a los inversores minoristas como racionalmente pasivos ante poder corporativo debido al mayor costo de la recopilación de información individual y su incapacidad para coordinarse colectivamente. el cambio comienza en [105], junto con estudios recientes que comienzan a explorar cómo el trading pasivo influye en la eficiencia de precios [106].

[107] y [108], entre diversos trabajos revisados en la literatura señalan que los ETF ofrecen liquidez elevada, menores costos de transacción y eficiencia en el"discovery"de precios, atributos que los convierten en instrumentos idóneos para trading de acciones. Un análisis exhaustivo destaca cómo los ETF, al seguir índices, permiten operar canastas de activos con rapidez, eficiencia y menores distorsiones de precio, mayores niveles de participación de ETF en acciones individuales se asocian con una mejor liquidez (menores costos de selección adversa y mayor resiliencia de precios), aunque esta ventaja puede disminuir o revertirse en momentos de tensión de mercado.

Queda en evidencia la necesidad de integrar estrategias indexadas dentro de un marco de aprendizaje reforzado profundo multiagente, segmentando el universo de activos en subíndices por área, esto permite que cada agente aprenda políticas de asignación especializadas en un espacio de estados reducido, mejorando la estabilidad y velocidad de convergencia del entrenamiento [109, 77].Estos artículos profundizan en la fusión de dos estrategias de trading financiero consolidadas, el seguro de cartera de proporción constante (CPPI) y la protección de cartera invariante en el tiempo (TIPP), con el marco de gradiente de política determinista profunda multiagente (MADDPG). Como resultado, se presenta dos nuevos métodos de AR multiagente (MARL): CPPI-MADDPG y TIPP-MADDPG, diseñados para explorar el trading estratégico en mercados cuantitativos. En [110] se configuran entornos virtuales con conjuntos de datos bursátiles, se entrena a los agentes de trading con redes neuronales y se analiza el rendimiento de las operaciones mediante backtesting exhaustivo. Además, incorpora importantes restricciones de trading, como el coste de las transacciones, la liquidez del mercado y el grado de aversión al riesgo del inversor.

Cada agente optimiza su portafolio mediante recompensas locales basadas en métricas de rendimiento y volatilidad de su índice, mientras se coordinan para la diversificación y busqueda de rentabilidades altas y estables[111, 112].

En el estudio de John Tidwell[113], se aborda el desafío de la negociación automatizada de acciones, donde los métodos tradicionales y el aprendizaje por refuerzo directo (RL) presentan dificultades con el ruido del mercado, la complejidad y la generalización, la solución propuesta es un marco integrado de aprendizaje profundo que combina una Red Neuronal Convolucional (CNN) para identificar patrones en indicadores técnicos con formato de imagen, una red de Memoria a Largo Plazo (LSTM) para capturar dependencias temporales tanto en el historial de precios como en los indicadores técnicos, y un agente de Red Q Profunda (DQN) que aprende la política de negociación óptima (compra, venta, retención) basándose en las características extraídas por la CNN y la LSTM.

Los índices proporcionan señales de mercado menos ruidosas y más representativas del comportamiento agregado, lo que redunda en políticas de trading más robustas y fáciles de interpretar [78].

En[114] el enfoque combina la información histórica de las ETFs con indicadores técnicos (por ejemplo, medias móviles, RSI, MACD) e información macroeconómica (por ejemplo, crecimiento del PIB, tasas de inflación) para ofrecer una perspectiva detallada de la dinámica del mercado. A través de estos datos, un modelo DRL de ensamble tiene como objetivo tomar decisiones de trading más adaptativas e informadas. El marco está diseñado para integrar los puntos fuertes de distintos algoritmos de aprendizaje reforzado profundo, para mejorar la robustez de las decisiones y minimizar el riesgo de sobreajuste. Una vez implementado realiza un proceso de evaluación para ver el rendimiento de la cartera y comparar el modelo ensamblado contra los modelos individuales. Luego,se tiene la hipótesis de que un enfoque MARL considerando índices (ETFs) en las áreas de segmentación de acuerdo al mercado chileno (sistema de AFP y gestoras de fondo) combinará la escalabilidad, liquidez y diversificación inherentes a estos activos con la adaptabilidad y optimización dinámica del aprendizaje reforzado multiagente.

5.1.1. Análisis matricial de activos de la AFP

En la Figura 5.3, la matriz de correlación de todos los activos visualizada con una escala de colores que varía de azul (correlaciones negativas) a rojo (correlaciones positivas), con valores intermedios en tonos de blanco y rosa según la intensidad facilita la identificación de relaciones entre los activos. Esto permite observar preliminarmente qué pares de activos tienden a moverse de manera similar o inversa, complementando el análisis de datos realizado en el capítulo 3 y por otra parte, fundamentando la necesidad del modelo Multiagente.

Al observar la matriz de correlación de la Figura 5.3 para todos los activos de renta fija y variable de las diferentes áreas, es posible identificar diversos patrones y puntos clave.

1. **Correlaciones Altas (Positivas)**:

   - **Bonos de empresas y bonos de gobiernos (Duraciones similares)**: Existen fuertes correlaciones entre los bonos de empresas y de gobiernos en duraciones similares, como los bonos de empresas de 5 años y los bonos de gobiernos de 4 años. Esto tiene sentido, ya que los activos de renta fija con duraciones cercanas y del mismo país tienden a moverse de forma similar en respuesta a factores macroeconómicos como la inflación y las tasas de interés.

   - **SP 500 y Mercados Globales (Europa, Japón, Emergentes, Asia)**: Se observa una alta correlación entre el índice SP 500 y otros mercados de renta variable internacional, especialmente en Europa y Japón. Esto es consistente con la interdependencia entre los mercados globales, donde los grandes índices tienden a tener movimientos sincronizados debido a la globalización y a la influencia de eventos económicos que afectan a múltiples países.

     Entre las empresas chilenas (SQM, BSANTANDER, FALABELLA, CENCOSUD), se encuentran algunas correlaciones positivas, especialmente entre aquellas del mismo sector o sectores relacionados (por ejemplo, entre bancos y minoristas). Esto refleja que factores económicos y financieros comunes en Chile impactan a las empresas locales de manera similar.

2. **Correlaciones Bajas (Negativas)**:

   - **Activos de renta fija y renta variable**: Las correlaciones entre bonos gubernamentales de largo plazo y el índice SP 500, son bajas o moderadamente negativas en algunos casos. Esto es típico, ya que los bonos suelen comportarse de manera opuesta a la renta variable en momentos de volatilidad, sirviendo como refugio seguro en mercados inciertos.

   - **Activos de largo plazo vs. corto plazo**: Los bonos de gobiernos con duraciones largas (16 UF) tienen una correlación más baja con activos de duración más corta, reflejando una menor sensibilidad a cambios de corto plazo en las tasas de interés. Los bonos a largo plazo se ven más afectados por la expectativa futura de inflación y otros factores de largo plazo.

3. **Aspectos Inusuales**

   - **Correlación entre bonos del gobierno y acciones chilenas**: La correlación positiva entre bonos de gobiernos y ciertas acciones específicas como CENCOSUD y BSANTANDER podría parecer inesperada, ya que en teoría, los bonos y las acciones tienen distintos factores de riesgo. Esto podría deberse a factores

específicos de la economía chilena o a movimientos de política económica que afectan tanto a los bonos como a algunas empresas chilenas.

- **Relación entre activos de mercados emergentes y algunos bonos**: Se observa que los bonos de gobiernos de duración intermedia (por ejemplo, 8 UF) tienen una correlación positiva con los mercados emergentes. Esto puede ser una rareza, ya que estos activos están expuestos a riesgos diferentes. Sin embargo, puede tener sentido en el contexto de flujos de capital hacia economías emergentes cuando los bonos chilenos también se perciben como activos de mayor rendimiento.

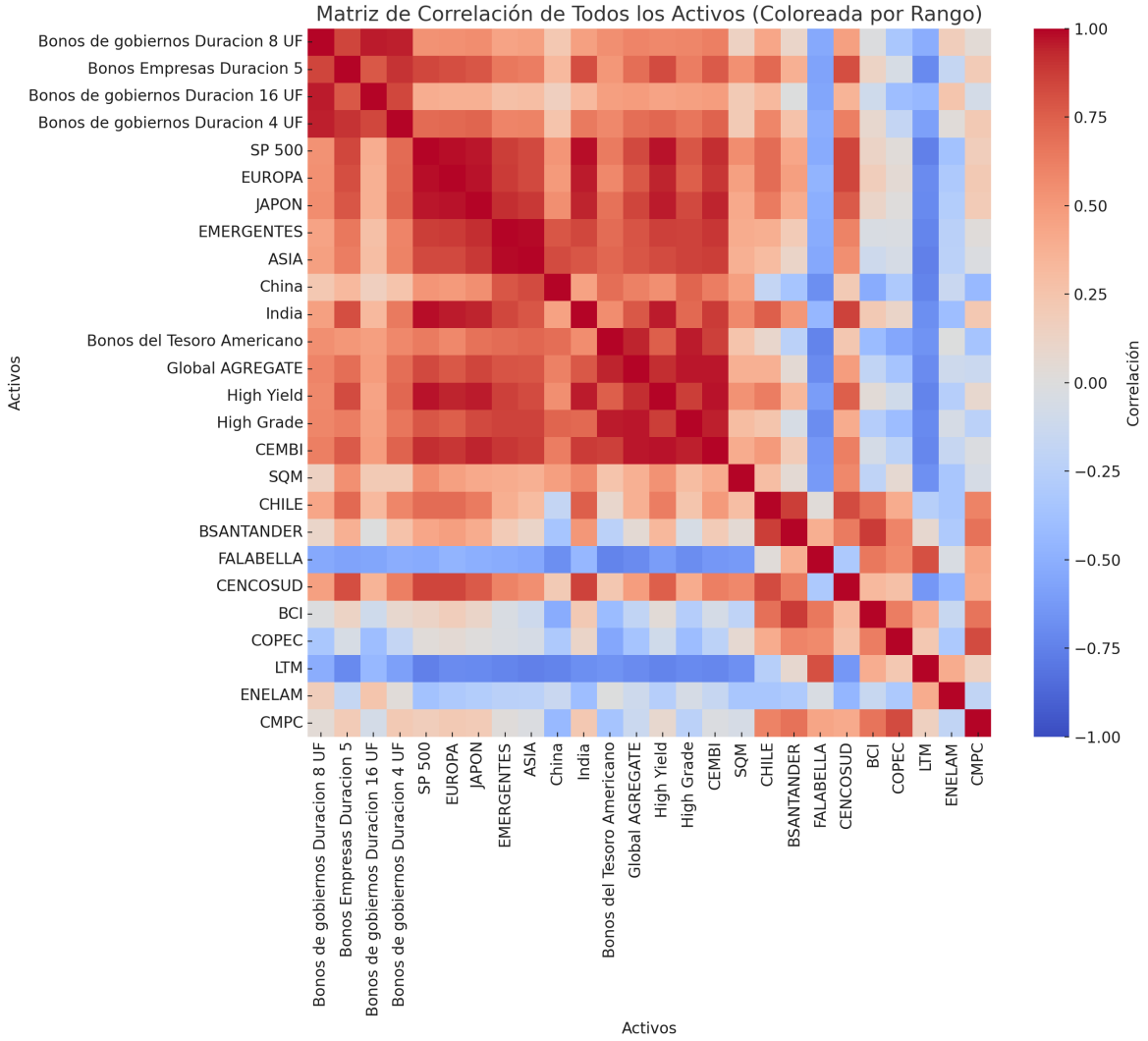

Figura 5.3: Matriz de Correlación.

5.2. Optimización multicriterio

Los resultados obtenidos de los diferentes modelos detallados en el capítulo 3, item 3.2 específicamente, en el espacio de riesgo y retorno muestran tres grupos, como se puede ver en la Figura 5.4.

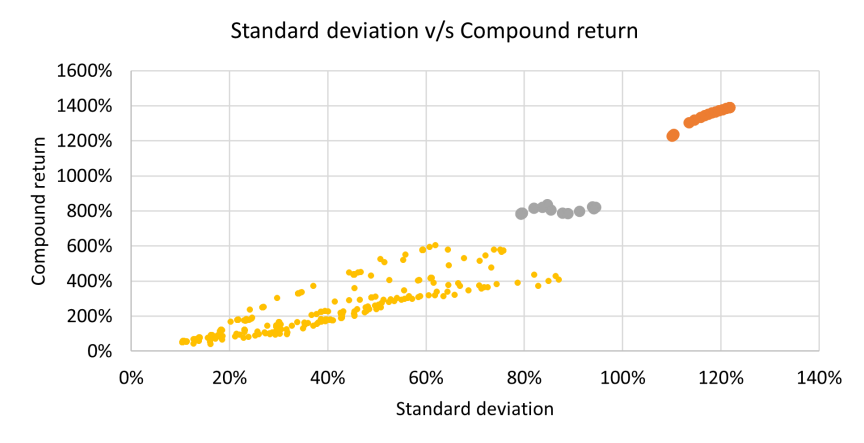

Figura 5.4: Varianza vs Retorno.

Esta figura muestra la desviación estándar de los portafolios frente a su retorno en esta evaluación de 20 años. Además, muestra el rendimiento compuesto en un período (acumulado o compuesto), identificando tres grupos según el nivel de rendimiento. El primer grupo con los rendimientos más altos, entre 1200 % y 1400 % o entre 1.11 % a 1.04 % en términos mensuales, correspondiente al 6.2 % (rojo) de los portafolios totales; un segundo grupo tiene rendimientos que oscilan alrededor del 800 % (equivalente a un rendimiento mensual de 0.87 %), también con un 6.2 % (gris) de los portafolios totales. Además de un tercer grupo, con el mayor número de elementos y rendimientos oscilando alrededor del 400 % (equivalente a un rendimiento mensual de 0.58 %) que corresponde al 87 % (amarillo), el cual se considera más representativo del total.

El primer grupo muestra una baja entropía, alrededor de 0.13, dominado por los portafolios de los modelos 2 y 5. El segundo grupo tiene una entropía de alrededor de 0.59, donde los portafolios de los modelos 2 y 5 dominan en cantidad. Con un comportamiento similar, el tercer grupo tiene una entropía de 0.83, que es alta en comparación con los otros modelos. Además, es importante notar que los portafolios dominantes en este grupo pertenecen al Modelo 7.

En la Figura 5.5, se presenta el portafolio resultante del modelo 7 ($\alpha = 0.3$, $\beta = 0.9$), en términos de retorno e índice de Sharpe, situado en el cuadrante superior derecho y demostrando un rendimiento superior al promedio tanto en retorno como en eficiencia. Además, destaca por su posición relativa a otros portafolios en el mismo cuadrante.

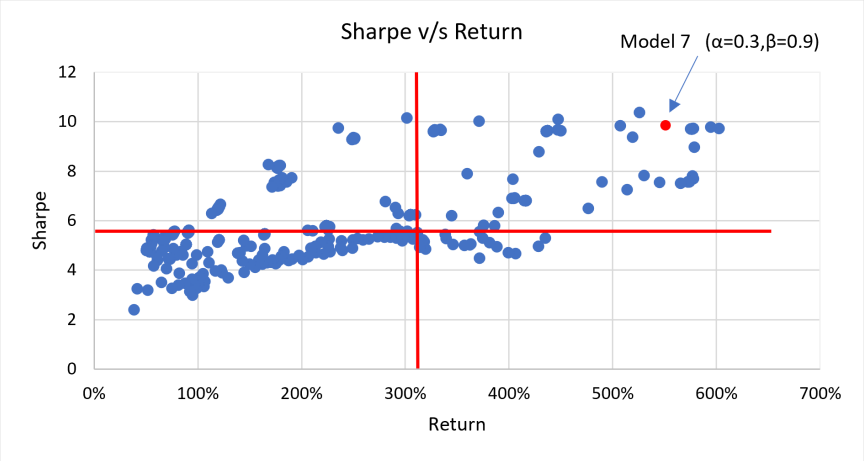

Figura 5.5: Retorno e índice de Sharpe.

Por otro lado, la Figura 5.6 ilustra que el modelo 7 se encuentra en la frontera eficiente al considerar los modelos en términos de minimizar el riesgo y maximizar los retornos. Es notable que el modelo 7, con $\alpha = 0{,}3$ y $\beta = 0{,}9$, exhibe uno de los mayores retornos y es uno de los más eficientes entre los portafolios derivados de diferentes modelos.

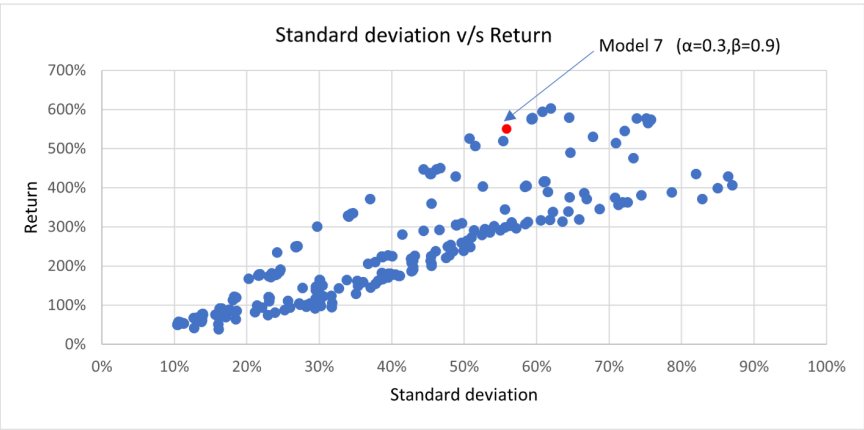

Figura 5.6: Retorno vs. Desviación estándar.

Además, en la Figura 5.7, se presenta una comparación del rendimiento de los modelos en los nueve escenarios propuestos, considerando sus diferencias en los valores de sus atributos (retorno, varianza y entropía).

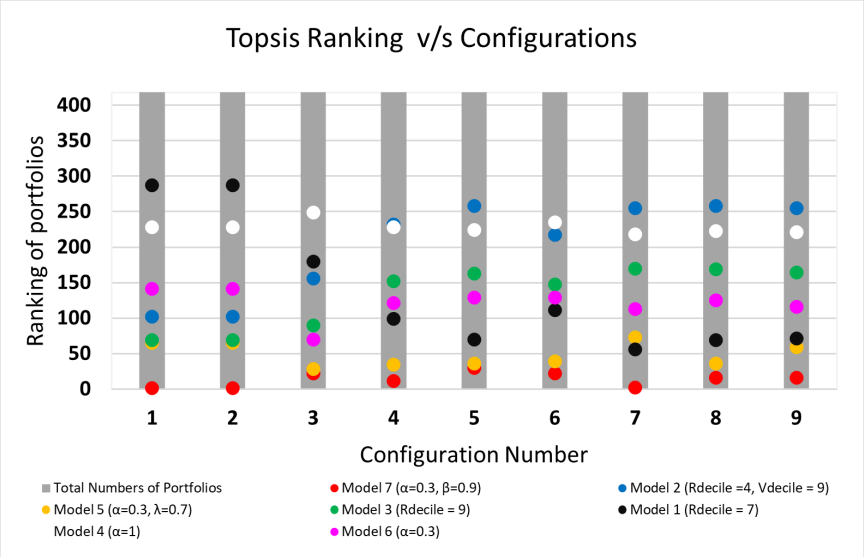

Figura 5.7: Clasificación de Modelos versus escenarios TOPSIS.

El propósito del análisis TOPSIS predicho en el gráfico es mostrar la posición de clasificación de un portafolio frente al resto de los portafolios en cada configuración. La longitud de la barra representa el número de modelos analizados, y la posición del punto en la barra indica qué tan bueno es el resultado. En este contexto, es posible notar que cuanto más bajo esté en la barra, mejor es el resultado.

Además, los modelos 5 y 7 exhiben un rendimiento superior en comparación con los otros modelos en diferentes configuraciones de TOPSIS. Del mismo modo, el modelo 5 se desempeña bien y es comparable al modelo 7, sin embargo, no supera al modelo 7. Por otro lado, los modelos 2 y 4 caen en la mitad superior del gráfico de TOPSIS. En particular, el modelo 4 se encuentra consistentemente en torno al medio de la clasificación en todos los escenarios.

En configuraciones donde el retorno es relevante, el modelo 2 es el peor, sin embargo, esta situación es opuesta al modelo 4 cuando la importancia del retorno es baja. Además, este modelo no se desempeña bien, pero el modelo 1 es el peor. Es relevante notar que el modelo 4 tiene un comportamiento estable, lo que significa que sus clasificaciones son similares entre escenarios. Si bien el modelo 6 es mejor que el 3, su comportamiento es similar al del 3, que se desempeña mejor que los modelos en la mitad del ranking.

Finalmente, en términos generales, se puede observar que el modelo 7 consistentemente exhibe un rendimiento superior en comparación con los otros modelos (obteniendo la primera posición en la clasificación) en diferentes configuraciones de las variables de decisión de TOPSIS. Por lo tanto, puede considerarse el mejor modelo en diferentes escenarios.

Los resultados del análisis sugieren que, para los datos y la metodología considerados, los modelos que incorporan medidas difusas y medidas de entropía difusa en entornos multiobjetivo, como los modelos 5, 6 y 7, demuestran un rendimiento superior en comparación con los modelos 1, 2 y 3. Además, es importante señalar que el modelo 4 no tiene un buen desempeño en comparación con los otros modelos difusos; cabe destacar que este modelo no cuenta con la función de pertenencia entrópica. En opinión personal, las medidas difusas proporcionan la flexibilidad para manejar intervalos e incertidumbres, y la entropía difusa maximiza la ganancia de información, tanto para el retorno como para la varianza, reforzando aún más la flexibilidad de los modelos.

Del mismo modo, en todo el rango de resultados, hay una notable dispersión en el rendimiento del portafolio, medida a través de los retornos y varianzas reales. Desde la perspectiva personal, esta dispersión puede atribuirse a dos factores principales:

1. **Uso de diferentes límites de varianza**: No solo en el modelo base de Markowitz, sino también en todos los demás modelos que incorporan la maximización de la entropía en sus versiones con restricción ϵ.

2. **Convergencia de modelos**: El hecho de que los resultados de los modelos de entropía difusa coinciden con los resultados de Markowitz para varianzas bajas debido al diseño inherente del modelo, que busca minimizar la varianza mientras maximiza el retorno.

En el caso de los modelos multiobjetivo 4 a 7, las varianzas realizadas se concentran en el rango medio, esto se debe a que se utilizan funciones objetivo para maximizar las funciones de pertenencia difusas de los parámetros. Para ambos casos considerados, λ para las instancias difusas y α y β para los casos de entropía difusa.

Por otro lado, el modelo 5 exhibe los mejores resultados en la clasificación cuando $\alpha = 0{,}3$ y los valores de λ cubren el intervalo [0 -1], lo que indica que la varianza está modulada por λ. Del mismo modo, los modelos 5, 6 y 7 muestran que la varianza real tiende a concentrarse en el rango medio, resultado de optimizar las funciones de pertenencia de los parámetros α y β.

Finalmente, el modelo 7, con $\alpha = 0{,}3$ y $\beta = 0{,}9$, logra una clasificación superior en varias configuraciones con diferentes ponderaciones de atributos. Esto indica la efectividad de los enfoques de entropía difusa para la optimización de retorno y varianza en nuestros datos.

5.3. Aprendizaje reforzado multiagente

En esta sección se presentan y analizan los resultados correspondientes a la selección de portafolios para las cuatro áreas consideradas: Renta Fija Nacional (RFN), Renta Variable Nacional (RVN), Renta Fija Internacional (RFI) y Renta Variable Internacional (RVI).

Para cada área, la evaluación del desempeño de las estrategias implementadas se realiza a través de distintas representaciones gráficas que permiten comprender la dinámica y evolución del modelo en distintos niveles. En particular, se incluyen: (i) gráficos de pesos asignados a los activos en función de las fechas, que ilustran la evolución y adaptación de la asignación de cartera durante el proceso de entrenamiento; (ii) gráficos de retorno acumulado en función de las fechas, que muestran la progresión del rendimiento a lo largo del tiempo; (iii) gráficos de retorno acumulado por etapas, que evidencian el progreso y la convergencia del modelo durante el proceso de aprendizaje; y (iv) gráficos de retorno acumulado en la fase de testeo, que permiten evaluar la capacidad de generalización y que tan robusta han sido las políticas aprendidas en datos no vistos. Estas visualizaciones en conjunto con las métricas de retorno acumulado, índice de Sharpe y Máximo Drawdown constituyen una herramienta integral para el análisis detallado y comparativo del desempeño en cada área, descritas a continuación facilitarán la identificación de fortalezas y posibles áreas de mejora.

- **Gráfico de asignación de pesos para fechas de entrenamiento**: Muestra la dinámica de asignación a cada activo en el portafolio a lo largo del proceso de entrenamiento, representado en función de las fechas. Cada línea o área representa la proporción del capital asignada a un activo específico en un momento dado.

  En las etapas iniciales, es común observar asignaciones relativamente uniformes o incluso aleatorias, ya que el modelo aún no ha aprendido las características óptimas de cada activo ni sus relaciones de riesgo-retorno. Durante este período, los pesos pueden fluctuar considerablemente, reflejando la exploración del espacio de soluciones.

  A medida que avanza el entrenamiento, el modelo comienza a identificar patrones y a ajustar la asignación de manera más coherente, privilegiando activos que contribuyen favorablemente al rendimiento del portafolio y reduciendo la exposición a activos con mayor riesgo o menor rentabilidad esperada. Este proceso se refleja en líneas más estables y segmentadas, donde ciertos activos predominan en determinados intervalos temporales.

  La evolución de los pesos responde a cambios en el entorno o en la información de

mercado, evidenciando la capacidad del modelo para adaptarse dinámicamente a condiciones variables y re-equilibrar la cartera de forma eficiente.

EL gráfico ilustra cómo el aprendizaje reforzado va construyendo una política de asignación que busca maximizar la recompensa acumulada ajustando progresivamente la distribución del capital entre los activos disponibles, con un balance entre exploración inicial y explotación madura hacia la convergencia.

- **Gráfico de retorno acumulado por fechas**: Refleja la evolución del desempeño del modelo conforme avanza su proceso de aprendizaje. Dado que cada etapa del entrenamiento implica una actualización de la política basada en una pasada completa sobre los datos históricos, la curva acumulada muestra cómo la política mejora progresivamente. En las fechas iniciales, el retorno es resultado de una política poco entrenada, mientras que en las fechas más recientes se evidencia un rendimiento superior y más estable, resultado de la última etapa del entrenamiento. De esta forma, el gráfico integra de manera implícita el progreso interno del modelo y su desempeño práctico en el mercado real.

- **Gráfico de retorno acumulado por etapas**: Muestra la evolución del rendimiento promedio del modelo a lo largo del proceso de entrenamiento, donde cada etapa representa una pasada completa del agente sobre toda la base de datos de entrenamiento. En este contexto, el eje horizontal indica el número de etapas, mientras que el eje vertical refleja el retorno acumulado promedio obtenido al final de cada etapa.

  En las etapas iniciales, es común observar una alta variabilidad y retornos relativamente bajos, ya que el agente comienza con políticas no optimizadas o casi aleatorias. Conforme avanza el entrenamiento, la política se va refinando mediante la interacción continua con el entorno, lo que se traduce en una mejora progresiva del retorno acumulado. Esta tendencia ascendente refleja la capacidad del agente para aprender y ajustar su comportamiento para maximizar la recompensa.

  Eventualmente, el gráfico exhibe una tendencia a la estabilización o convergencia, donde las mejoras en el retorno acumulado se vuelven marginales. Esta fase indica que el agente ha alcanzado un nivel óptimo o cercano al óptimo en su política, y que continúa aplicándola con rendimiento consistente. En algunos casos, pueden observarse pequeñas fluctuaciones o retrocesos temporales, los cuales son naturales debido al ruido inherente en los datos y al proceso de exploración-explotación. Este gráfico es fundamental para evaluar la eficiencia y estabilidad del proceso de entrenamiento en modelos de aprendizaje reforzado, ya que permite identificar la etapa a partir de la cual el agente muestra un desempeño satisfactorio y evitar el sobre-entrenamiento o estancamiento prematuro, de acuerdo a lo establecido en el

capítulo 4, respecto de la metodología.

- **Gráfico de Retorno acumulado en fase de testeo**: El gráfico de retorno acumulado en la fase de testeo es fundamental para evaluar la capacidad de generalización y robustez de la estrategia de inversión. Al partir desde una base 1, se puede observar la evolución del capital invertido bajo la política derivada del entrenamiento, permitiendo identificar periodos de crecimiento, estabilidad o caída.

  Un desempeño sólido en esta etapa indica que la política ha aprendido patrones relevantes y que es capaz de adaptarse a condiciones de mercado no vistas, minimizando riesgos y maximizando retornos.

  Las fluctuaciones o caídas, por otro lado, reflejan la volatilidad inherente del mercado y posibles limitaciones del modelo para adaptarse a escenarios imprevistos.

  El análisis gráfico de entrenamiento y testeo se realiza junto a tres métricas cuantitativas, el retorno acumulado final, el índice de Sharpe y el máximo drawdown brindando una visión integral del valor práctico y confiabilidad de cada una de las estrategias evaluadas.

5.3.1. Renta fija nacional (RFN)

Para comenzar y de acuerdo al análisis cuantitativo de los datos de la serie de precios diarios de Renta Fija Nacional efectuada en el capítulo 3, los activos analizados presentan retornos promedio positivos, con los bonos de gobierno a 16 UF y los bonos de empresas de duración 5 liderando con medias semanales de 0.147 % y 0.131 % respectivamente, mientras que los bonos de gobierno a 4 UF registran el menor rendimiento medio (0.078 %). Desde la perspectiva del riesgo, la volatilidad histórica semanal es más elevada en la curva 16 UF (2.49 %) y más contenida en el tramo más corto de 4 UF (0.59 %), lo que evidencia una prima de riesgo creciente con el plazo. Las distribuciones de retorno muestran asimetrías negativas moderadas indicando caídas ocasionales más extremas que las subidas, y curtosis superior a 3 en todos los casos, lo que sugiere colas más pesadas y eventos extremos en este mercado. Estos hallazgos sirven de base para mejorarlas métricas mediante la modelación y fundamentan el diseño de la política DDPG, pues una correcta asignación de pesos debe balancear la maximización del retorno local con la mitigación de la mayor volatilidad y el riesgo de eventos extremos en los tramos más largos.

En lo que respecta a la construcción del modelo de aprendizaje por refuerzo profundo, con un agente MLP, un algoritmo DDPG, y una función de recompensa que incorpora de manera explícita el retorno absoluto, junto con el índice Sharpe local y las penalizaciones por costos de transacción, cambios en duración y variaciones en convexidad, permite

corregir el sesgo que surge al optimizar únicamente métricas relativas de riesgo en activos de muy baja volatilidad, cuya estabilidad tiende a generar índices de Sharpe elevados pero retornos absolutos reducidos. Al añadir el término de retorno observado como componente positivo de la recompensa, el algoritmo favorece asignaciones que no solo mantengan un perfil ajustado por riesgo eficiente, sino que también maximicen la rentabilidad absoluta. De esta forma, la política aprendida alcanza un balance técnico entre estabilidad y generación de retorno, condición fundamental para una gestión óptima de portafolios de renta fija nacional.

Este proceso busca extraer las características más deseables en un portafolio, altos retornos y alta eficiencia al adaptar las decisiones de inversión de manera continua.

En la Figura 5.8 y a la Figura 5.9 se presenta una exposición integrada de los resultados obtenidos para la evolución de los pesos dinámicos en el portafolio de Renta Fija Nacional (RFN) bajo las cuatro configuraciones de función de recompensa propuestas. En cada caso, los pesos están normalizados de modo que suman 1 en cada instante.

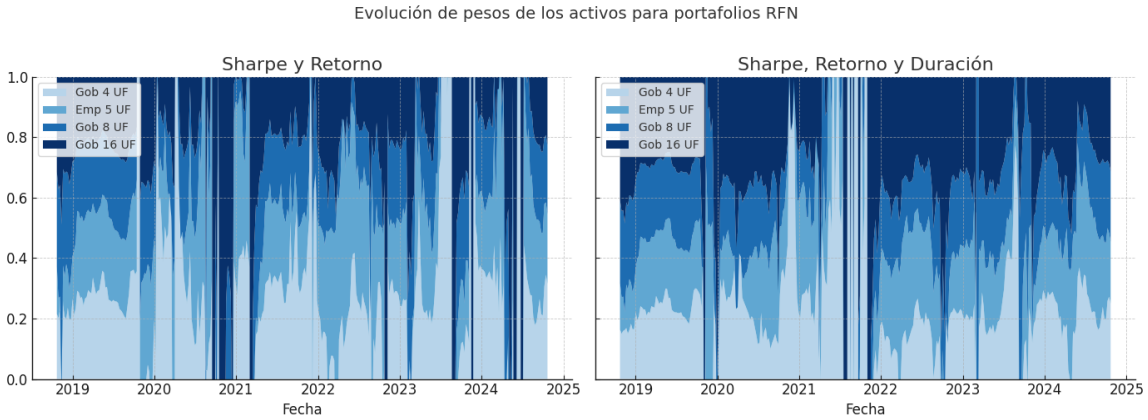

Figura 5.8: Escenarios del Portafolio RFN.

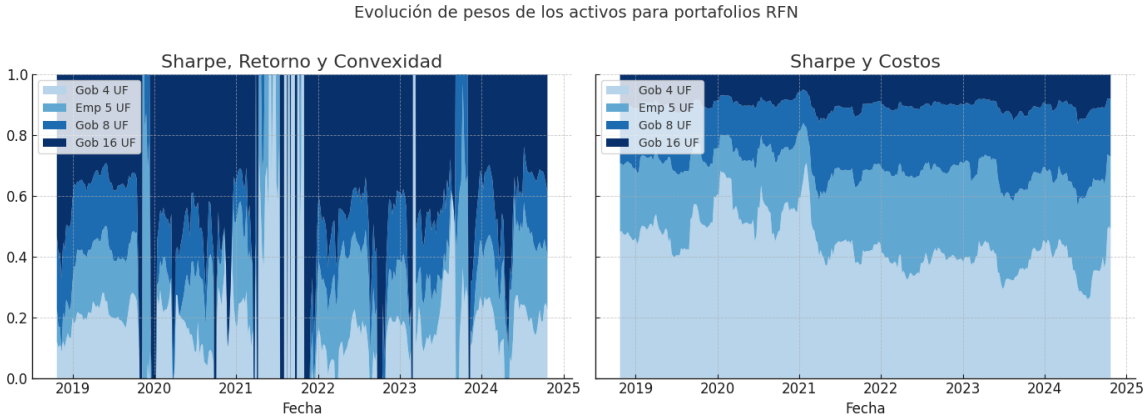

Figura 5.9: Otros escenarios del Portafolio RFN.

1. **Escenario "Sharpe y Retorno"**: Al combinar únicamente el índice de Sharpe local y el retorno , el agente asigna en promedio un 20.07 % del capital a los bonos de gobierno de duración 16 años (desvío estándar de 5.63 %) y un 25.45 % a los de duración 8 años (desvío estándar de 1.93 %), mientras que los bonos corporativos de duración 5 años obtienen un 25.34 % (desvío estándar de 3.64 %) y los de gobierno de duración 4 un 29.14 % (desvío estándar de 5.83 %). Esta distribución dinámica de pesos refleja concentración en los activos de duración más baja, lo cual va en linea con la función de recompensa que incorpora al índice de Sharpe como componente. Es relevante destacar que los activos de más baja duración presentan un mayor valor de dicho índice con respecto a los activos de mayor duración.

2. **Escenario "Sharpe, Retorno y Duración"**: Al incorporar además un término de penalización en los cambios de duración, el portafolio gira decididamente hacia vencimientos más largos. El tramo de gobierno de duración 16 concentra en promedio el 30.99 % del peso (4.70 % de variabilidad), seguido por los bonos de duración 8 (22.70 % ± 4.26 %) y los corporativos de duración 5 (24.53 % ± 2.87 %), mientras que el bono de gobierno de duración 4 retrocede al 21.78 % ± 3.18 %. La inclusión de duración empuja al agente a seleccionar activos que, aun con retornos ajustados y Sharpe adecuados, ofrezcan una sensibilidad al tipo de interés consistente con el trade-off definido.

3. **Escenario "Sharpe, Retorno y Convexidad"**: Al premiar la convexidad, el agente asigna un mayor valor al peso del activo de mayor duración, de tal manera que, las asignaciones corresponden a un 45.41 % ± 2.62 p.p. al tramo de duración 16, con pesos medios del 17.94 % ± 2.84 p.p. en duración 8 , 17.31 % ± 2.83 p.p. en corporativos de duración 5 y 19.34 % ± 1.93 p.p. en activos de duración 4. Al reforzar la convexidad, la política RL aprende a privilegiar instrumentos con mayores curvaturas en la relación precio–tasa, mejorando la protección ante movimientos extremos sin sacrificar excesivo rendimiento.

4. **Escenario "Sharpe y Costos"**: Finalmente, este escenario es más estable que los anteriores porque se penaliza el cambio de pesos de los activos habiendo una proporción mayoritaria de bonos de baja duración. El activo de duración 4 alcanza un peso medio del 42.42 % ± 3.88 %, seguido por 5 UF con 28.64 % ± 5.63 %, 8 UF con 19.55 % ± 1.85 % y 16 UF con apenas 9.39 % ± 1.34 %. Esta asignación evidencia la lógica de la función de recompensa, dado que busca maximizar el índice de Sharpe considerando penalizar los movimientos o cambios en los pesos del portafolio.

En conjunto, estos cuatro escenarios ilustran cómo, mediante aprendizaje reforzado, es posible dirigir la asignación de un portafolio de renta fija nacional hacia distintos objetivos,

como son: maximizar retornos ajustados y controlar sensibilidad a tasas, minimizar cambien en la convexidad o limitar costos todo ello con una única arquitectura RL (DDPG) que adapta su política en función de la forma de la función de recompensa. Las estadísticas de pesos medios y su dispersión confirman asignaciones estables y coherentes con las prioridades establecidas para cada configuración.

En la Figura 5.10, el portafolio final optimizado para RFN refleja un equilibrio entre los objetivos impuestos por cada una de los componentes de la función de recompensa.

Al entrenar un agente cuya función de recompensa integra de manera conjunta cinco componentes Sharpe local,retorno absoluto, costos de transacción, cambio de duración y convexidad, hemos obtenido una política que ajusta los pesos de los cuatro segmentos de mercado de manera continua y coherente con los costos de oportunidad asociados al mercado. Bajo esta configuración la asignación media se distribuye de modo equilibrado entre los extremos de la curva: aproximadamente 38.43 % en bonos de duración 16, 21.57 % para duración 8 y 19.95 % en corporativos de duracion, relegando el bono de gobierno de duración 4 a 20.05 %.

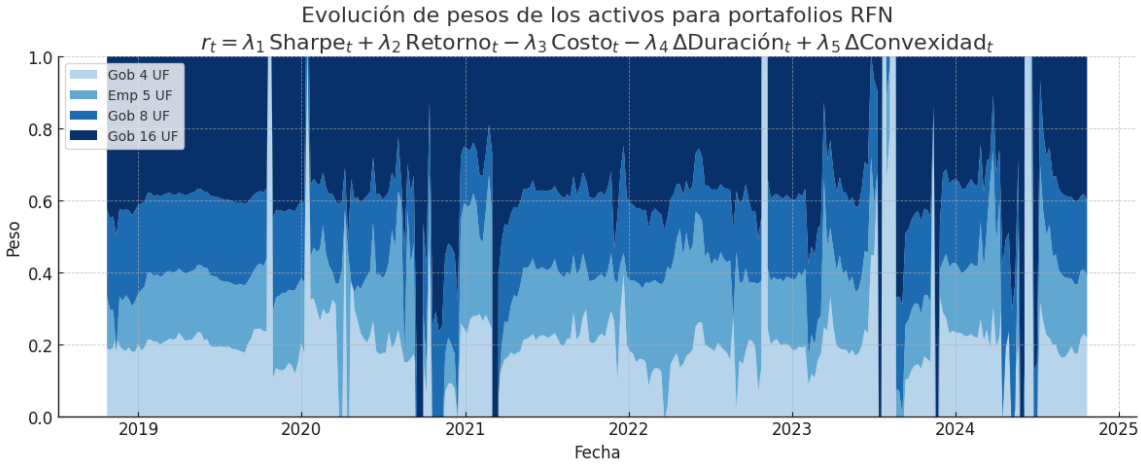

Figura 5.10: Portafolio RFN.

Esta configuración refleja el interés por instrumentos que combinan retorno ajustado por volatilidad sin perder el enfoque en la rentabilidad , tal como lo haría un gestor de inversiones. Al añadir la penalización sobre los cambios en la duración, se añade la estrategia de incrementar la exposición a vencimientos largos cuando se desea un mayor carry (rentabilidad que obtiene un inversor por simplemente mantener el bono en cartera durante un periodo dado, bajo la hipótesis de que las curvas de tipos de interés no se mueven) o se anticipa un descenso de tasas, manteniendo al mismo tiempo un nivel controlado de volatilidad. Por último, la convexidad agrega la preferencia por la curvatura de precio frente a cambios de tasa, asegurando una protección extra ante movimientos de

mercado bruscos sin sacrificar entonces el retorno. Al incluir un término que penaliza los costos asociados a rebalanceos siempre en combinación con el índice de Sharpe, el agente considera el impacto de las comisiones. De tal manera que el portafolio sujeto a la función de recompensa combinada, entrega un perfil de pesos dinámicos que se encuentra entre un modelo que considera como objetivos únicos el Sharpe y el Retorno, altamente volatil en la asignación de los pesos en el tiempo y el modelo que considera solo Sharpe y costos de transacción de baja volatilidad en la asignación de los pesos.

En particular la aplicación de DDPG a la asignación dinámica de un portafolio de renta fija nacional ha demostrado ser capaz de reproducir y, al mismo tiempo, enriquecer las prácticas clásicas de gestión de activos.

Durante la fase de entrenamiento el análisis del desempeño acumulado de distintas estrategias de construcción de portafolio sobre activos de renta fija nacional (RFN), en el periodo comprendido desde el 31 de julio de 2018 hasta el 22 de octubre de 2024 presentado en la figura 5.11 revela diferencias sustanciales en cuanto a la eficiencia de cada modelo. En la figura correspondiente se representan cuatro enfoques: un modelo de aprendizaje reforzado multiagente (MARL), un modelo uniagente de aprendizaje reforzado (RL Uniagente), una estrategia clásica de optimización media-varianza (Markowitz), y un portafolio con asignación equitativa de pesos (Equiweight). El retorno acumulado ha sido normalizado desde el valor 1 al inicio del entrenamiento, lo que permite una realizar una comparación directa entre modelos.

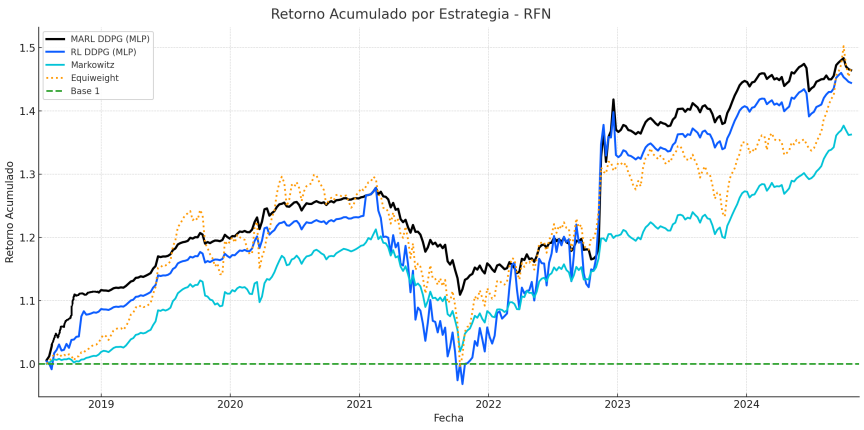

Figura 5.11: Portafolio MARL RFN (global) UNI RFN en entrenamiento.

Las curvas de retorno acumulado muestran que las estrategias basadas en aprendizaje reforzado (DDPG y MARL) superan consistentemente a las estrategias tradicionales (EW y MKW), la estrategia MARL alcanzó un retorno acumulado final aproximado de 1.48, es decir, una ganancia del 48 % sobre la inversión inicial, superando el 44 % de DDPG, el 38 % de MKW y el 43 % de EW. sin embargo las mayores diferencias se observan a tramos

durante el entrenamiento, más que en términos acumulados al final del período.

Durante la fase de entrenamiento (2018–2024), se observa que el modelo AR uniagente y el Modelo MARL mantienen una trayectoria ascendente sostenida y aunque la diferencia numérica al final del período equivale a aproximadamente el 5 % del Modelo MARL sobre el modelo uniagente , lo cual parece modesto en términos absolutos, pero tiene un impacto financiero significativo dada la naturaleza compuesta del retorno y el horizonte temporal considerado.

El gráfico muestra además que la curva MARL no solo supera en rendimiento sino que exhibe una convergencia más suave y estable, con menos oscilaciones pronunciadas que el RL uniagente. Esta estabilidad sugiere que el aprendizaje multiagente es capaz de optimizar la política de inversión de manera más robusta y consistente, evitando fluctuaciones bruscas que podrían traducirse en riesgos. La mayor estabilidad y rendimiento del MARL evidenciados en la gráfica reflejan que no hay estancamiento en óptimos locales subóptimos o sobreexplotación de estrategias no generalizables. La curva ascendente y menos volátil indica que los agentes coordinados logran políticas que maximizan el retorno esperado mientras controlan la variabilidad, clave para la gestión del riesgo.

En términos de riesgo, el análisis del índice de Sharpe y el máximo drawdown confirma la solides de MARL. Con un índice de Sharpe aproximado de 1.25, MARL ofrece la mejor compensación entre retorno y volatilidad, seguido por un modelo uniagente DDPG (MLP) con 1.10. Las estrategias clásicas de Markowitz (MKW) y equiweight (EW)presentaron Sharpes significativamente menores, 0.75 y 0.60 respectivamente, reflejando menor eficiencia. Por otra parte, MARL exhibió un máximo drawdown del -12 %, considerablemente menor que el -18 % de DDPG (MLP) y que el -16 % y -18 % de MKW y EW, respectivamente. Esta menor exposición a caídas severas resalta la capacidad del modelo MARL para proteger el capital en escenarios adversos. como se muestra en la Tabla 5.1

Cuadro 5.1: Resumen de métricas financieras para estrategias evaluadas

Estrategia	**Retorno acumulado**	**Sharpe ratio**	**Máximo Drawdown**
MARL	1.48	1.25	-0.12
DDPG (MLP)	1.43	1.10	-0.18
Markowitz (MKW)	1.38	0.75	-0.16
EquiWeight (EW)	1.44	0.60	-0.18

La jerarquía observada subraya el valor añadido del aprendizaje automático y, particularmente, del enfoque multiagente, que supera la rigidez de modelos tradicionales y se adapta mejor a las condiciones cambiantes del mercado.

La máxima caída (drawdown) para el modelo uniagente resulta (–18 %) frente al modelo MARL de orden de magnitud (–12 %). Estos valores se explican por las características propias de RFN, donde la autocorrelación positiva en horizontes cortos y los cambios de régimen de las tasas de interés favorecen una política global que capture la estructura autoregresiva de los retornos gracias a una mayor estabilidad que las acciones. El enfoque multiagente, al fragmentar la señal en cuatro subagentes especializados, puede introducir cierta complejidad y diluir la predictibilidad autoregresiva, lo que sería una desventaja del modelo MARL,punto que se podría evidenciar en alguna medida al comienzo de la serie estadística, desde inicios del 2021 hasta mediados del 2022 y de alguna manera osilatoria, este es un punto que debiera tenerse en consideración a la hora de la generación de modelos de construcción de portafolios.

Además, se muestra la evolución del retorno acumulado promedio del portafolio en RFN a lo largo de 350 etapas de entrenamiento, donde cada etapa representa una pasada completa del modelo sobre todos los datos disponibles. Se presentan dos curvas correspondientes a dos estrategias de aprendizaje reforzado: MARL y RL uniagente.

En las primeras etapas, ambas curvas inician en torno a la unidad, reflejando el capital inicial sin ganancias ni pérdidas. Durante este periodo inicial, se observa una mejora gradual en el retorno acumulado, aunque con fluctuaciones naturales debidas a la exploración del espacio de políticas y al ruido inherente del entorno financiero.

Conforme avanza el entrenamiento, las curvas muestran una tendencia ascendente más marcada, indicando que los modelos van aprendiendo políticas cada vez más eficientes para la asignación de activos y la maximización del retorno ajustado al riesgo. La persistencia de ciertas oscilaciones en etapas intermedias y avanzadas sugiere la complejidad del entorno y la adaptación constante del modelo a las condiciones cambiantes del mercado.

Finalmente, cerca de la etapa 350, las curvas comienzan a estabilizarse para ambos modelos, la gráfica particularmente volatil en torno a una curva ascendente muestra una necesidad continua de exploración y explotación con un claro acercamiento a la convergencia del aprendizaje con la existencia de fluctuaciones moderadas durante el proceso que refleja además la naturaleza estocástica del entorno financiero.

MARL exhibe un retorno acumulado ligeramente superior a RL uniagente, indicando una mejor capacidad de aprendizaje y gestión del portafolio en este contexto específico.

Estas observaciones permiten inferir que el proceso de entrenamiento es efectivo y que la metodología basada en aprendizaje reforzado mejora progresivamente el desempeño del portafolio.

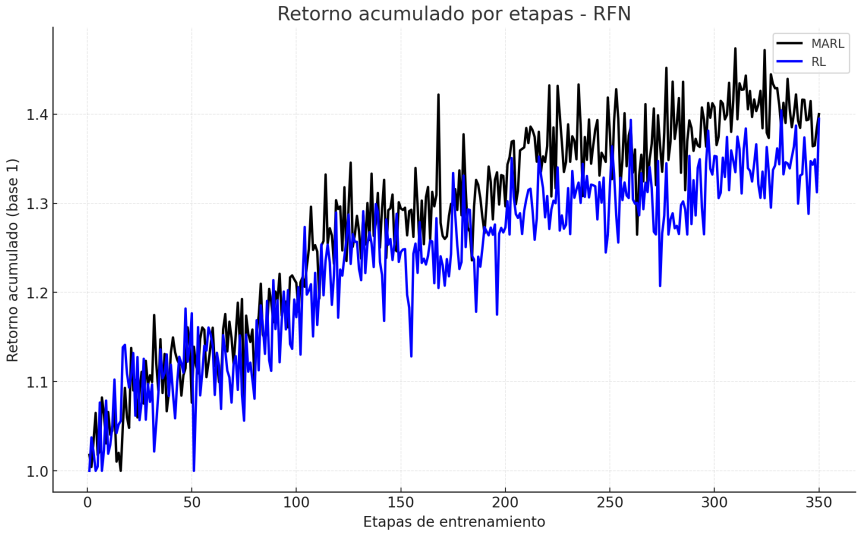

Figura 5.12: Portafolio MARL RFN (global) - UNI RFN en entrenamiento.

La fase de testeo considerada desde noviembre del año 2024 hasta mayo de 2025, evalúa el desempeño de los modelos bajo una política fija, entrenada previamente. En esta etapa, el retorno acumulado se reinicia desde 1, donde el testeo se concibe como un nuevo episodio que valida la robustez de la política aprendida. Desde el punto de vista teórico, esta práctica asegura que el rendimiento observado sea atribuible exclusivamente a la política implementada, sin arrastre de desempeño histórico.

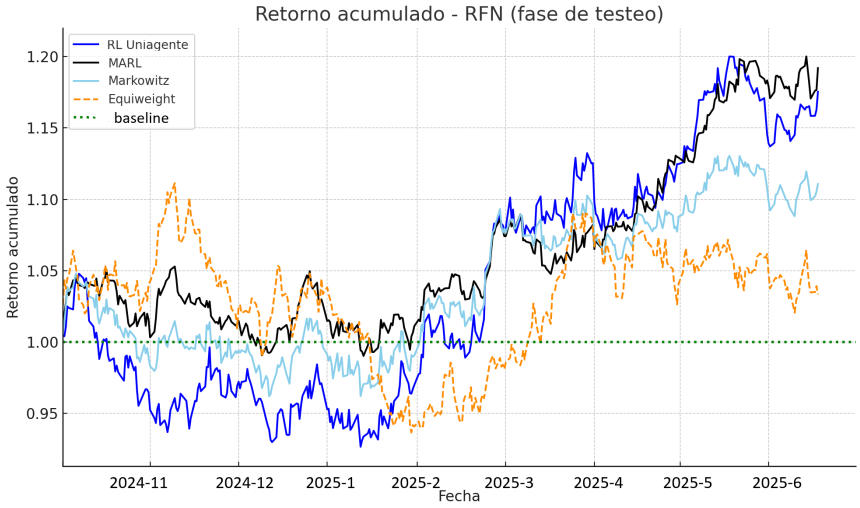

Figura 5.13: Portafolio MARL RFN (global) - UNI RFN en entrenamiento.

En la figura 5.13 se comparan las cuatro estrategias, durante los primeros meses de 2024, se observan oscilaciones menores sin una clara separación entre estrategias. Sin embargo, a partir del segundo semestre de 2024 se evidencia una diferenciación progresiva. El modelo MARL (representado por la curva negra) comienza a mostrar un rendimiento

creciente sostenido, alcanzando hacia mediados de 2025 un retorno acumulado cercano a 1.20. Esta trayectoria indica que la política aprendida por el coordinador y el agente único logra capturar señales efectivas del mercado durante el entrenamiento, aplicándolas de manera eficiente bajo condiciones no vistas.

El modelo RL, si bien mantiene una tendencia positiva, presenta un rendimiento algo inferior en la fase de testeo. Su retorno acumulado final bordea los 1.19, siendo equivalente al modelo MARL, pero demuestra mayor volatilidad y un máximo drawdown más profundo en 2024, evidenciando mayor riesgo durante períodos adversos. Esta diferencia, es relevante cuando va en línea con los resultados del entrenamiento, ya que sugiere que en este entorno específico, el mecanismo uniagente, es inestable. Sin embargo en términos de la coordinación multiagente también podría haber inducido restricciones que limitaron parcialmente la capacidad de adaptación frente a nuevas dinámicas.Se producen cruces temporales entre los modelos MARL y RL, sin embargo gracias al mejor manejo de la volatilidad y las caídas del modelo MARL, este último presenta mayor estabilidad (MDD=-0.05) de acuerdo a la Tabla 5.2.

Por otra parte, la estrategia Markowitz se posiciona sistemáticamente por debajo del modelo MARL. Tiene un Sharpe bajo de 0.3 y un drawdown considerable, lo que refleja su vulnerabilidad en mercados volátiles. Su pendiente se reduce progresivamente a partir del año 2025, reflejando una pérdida de efectividad frente a cambios estructurales del mercado. Finalmente, la estrategia Equiweight muestra un comportamiento prácticamente plano en comparación a los modelos comparativos, con retornos acumulados cercanos al nivel base, lo que confirma su carácter no adaptativo.

Desde el punto de vista técnico, se demuestra que el aprendizaje reforzado permite alcanzar rendimientos superiores a los métodos clásicos en fases fuera de muestra. En particular, el modelo MARL logra en este caso un mejor desempeño que RL, lo que destaca la importancia del equilibrio entre flexibilidad de la política y robustez estructural, clave para la toma de decisiones en contextos de asignación dinámica de activos.

Cuadro 5.2: Métricas de desempeño para estrategias RFN - fase de testeo

Estrategia	**Retorno Acumulado Final**	**Sharpe Ratio**	**Máximo Drawdown**
MARL	1.19	1.20	-0.05
RL Uniagente	1.18	0.90	-0.08
Markowitz	1.12	0.30	-0.12
Equiweight	1.04	0.50	-0.07

5.3.2. Renta fija internacional (RFI)

De acuerdo al Capítulo 4, sección 4.1, los datos de Renta Fija Internacional (RFI) reflejan una combinación de activos con diferentes niveles de riesgo, retorno y sensibilidad a las tasas de interés y el riesgo país. Los bonos del Tesoro Americano se presentan con el mayor promedio de precios y representan la opción más segura dentro del conjunto, sirviendo como activo refugio en momentos de crisis.

Respecto al Global Aggregate, se índica que este índice es una opción diversificada en sí misma, capturando el comportamiento del mercado global de renta fija. Su incorporación aporta diversificación internacional y exposición a distintas economías con menor riesgo que la renta variable. Del mismo modo, el índice High Yield refleja mayor riesgo crediticio en comparación con otros bonos de renta fija. Su inclusión en una cartera puede mejorar el rendimiento. El índice High Grade y High Yield, ofrece un punto intermedio entre rentabilidad y seguridad. Finalmente, CEMBI refleja por definición el comportamiento de la deuda corporativa en economías emergentes. Si bien su rentabilidad puede ser atractiva, es importante que su riesgo país y exposición a monedas extranjeras sean gestionados adecuadamente. En este contexto, al analizar la evolución de pesos obtenida por las distintas estrategías, los resultados de asignación de portafolios obtenidos en la Figura 5.14 deben ser coherentes con el análisis expuesto.

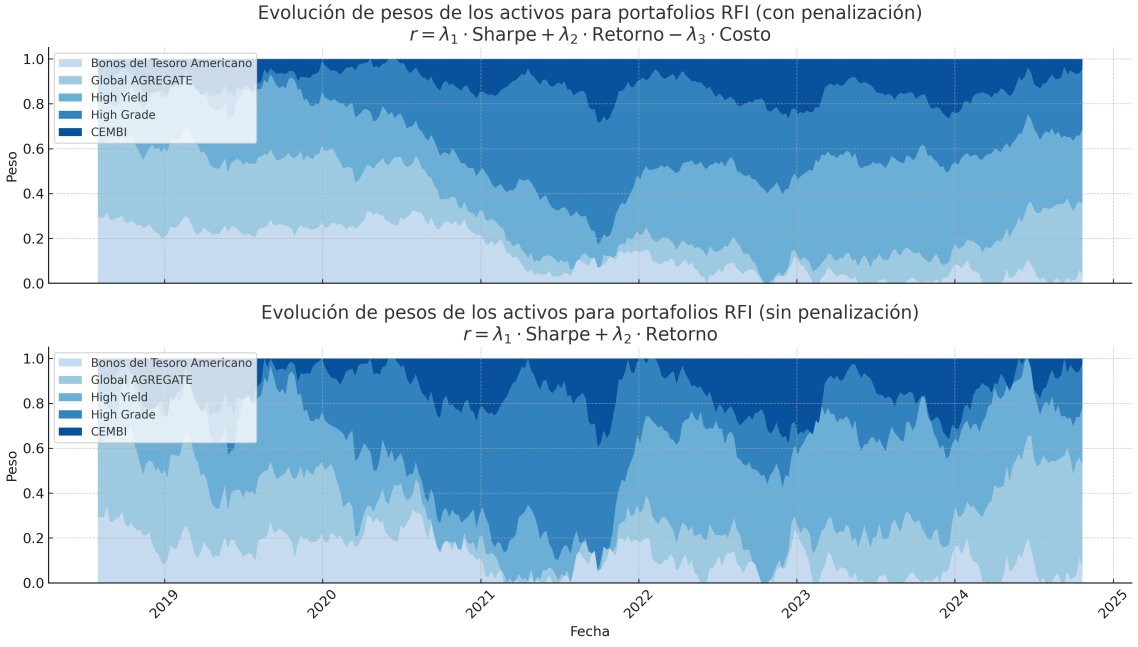

Figura 5.14: Evolución de los pesos del portafolio RFI MLP RL-PPO.

La inclusión o exclusión de la penalización por costos de transacción en la función de recompensa utilizada por el algoritmo PPO tiene implicancias relevantes tanto en el comportamiento dinámico de los pesos del portafolio como en su composición promedio.

Cuando el modelo considera explícitamente los costos de transacción (i.e., incluye un término negativo proporcional al tamaño del rebalanceo en la función de recompensa), la evolución de los pesos en el tiempo es marcadamente más suave. Se observa que el algoritmo evita realizar cambios bruscos en las asignaciones, y por tanto reduce la frecuencia de ajustes entre los activos del portafolio. Este comportamiento genera trayectorias más estables y conservadoras, compatibles con políticas de inversión que buscan minimizar la fricción operativa y preservar la eficiencia ante escenarios de alta rotación. En contraste, al remover dicha penalización, el algoritmo responde de forma mucho más agresiva a las señales de mercado: los cambios en los pesos son abruptos, con saltos significativos en la asignación entre activos incluso en intervalos breves. Esta estrategia refleja un enfoque oportunista y puramente rentista, en donde se privilegia la maximización del retorno y del ratio de Sharpe sin atender al impacto que conlleva la ejecución continua de rebalanceos.

Desde una perspectiva cuantitativa, estas diferencias se traducen en variaciones sustanciales en la asignación promedio de los activos. En el escenario con penalización por costos, el activo con mayor peso medio fue High Yield, con un 29.9 % del portafolio, seguido de High Grade con un 26.6 % y Global Aggregate con 17.5 %. Bonos del Tesoro Americano mantuvo una participación del 13.6 % y CEMBI un 12.4 %. En cambio, en el escenario sin penalización, High Yield permanece como el activo dominante con un promedio similar (29.4 %), pero el resto del portafolio se redistribuye con mayor libertad: Global Aggregate sube a 19.8 % y CEMBI también aumenta su participación promedio a 14.7 %, mientras que High Grade y Bonos del Tesoro Americano descienden a 24.3 % y 11.8 %, respectivamente.

Esta redistribución es coherente con lo que la teoría sugiere: al eliminar la penalización por costos de transacción, el algoritmo puede perseguir más activamente señales de retorno en activos como CEMBI o Aggregate sin estar limitado por el costo de mover capital entre ellos. Por el contrario, la versión penalizada del modelo adopta una política más prudente, en la que la estabilidad de los pesos y la minimización del rebalanceo adquieren un rol central. Así, el costo implícito de cambiar de posición genera un sesgo hacia configuraciones más persistentes, en las cuales ciertos activos defensivos, como los Bonos del Tesoro o los instrumentos investment grade, mantienen un lugar preferencial incluso en escenarios de mayor rendimiento en otras clases de activos.

En síntesis, la comparación de ambos enfoques permite ilustrar claramente cómo la incorporación de fricción financiera en la función de optimización no solo impacta la dinámica del portafolio, sino también la composición promedio de la cartera. Esta sensibilidad a la estructura de incentivos resalta la potencia y flexibilidad del aprendizaje por refuerzo para capturar múltiples objetivos financieros, y al mismo tiempo pone de relieve la necesidad de definir cuidadosamente la función de recompensa en función de las

restricciones reales del mercado y las metas del inversionista.

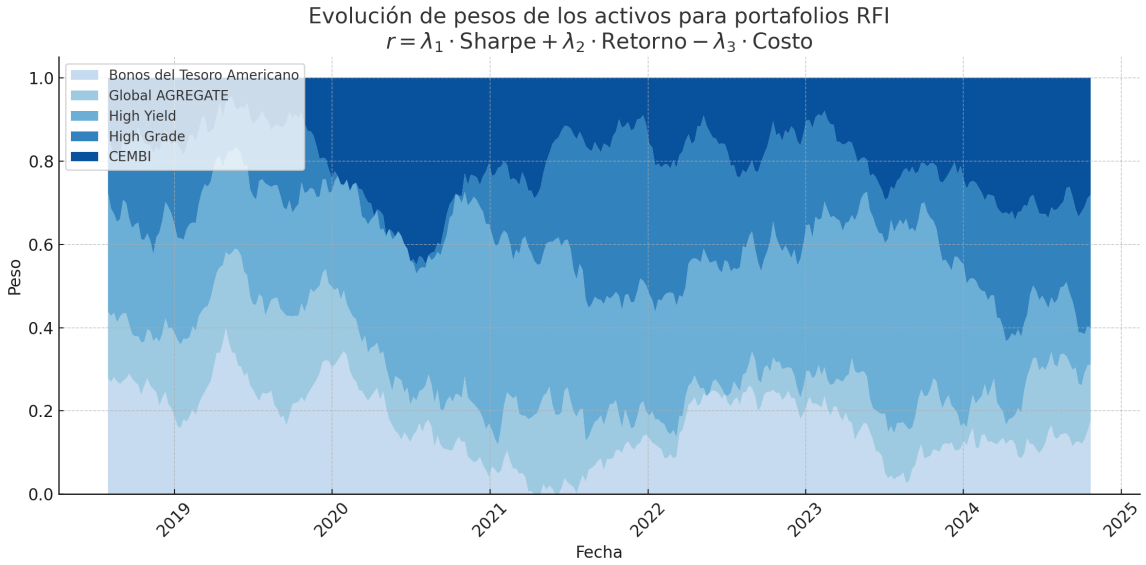

Figura 5.15: Evolución de los pesos del portafolio RFI en el tiempo.

En un entorno MARL, como se muestra en la Figura 5.15 la evolución de los pesos del portafolio asignados por el algoritmo, refleja de forma coherente los principios tanto del aprendizaje por refuerzo como de la teoría financiera moderna. El gráfico generado muestra la asignación dinámica de proporciones de inversión en cinco clases de activos: Bonos del Tesoro Americano, Global Aggregate, High Yield, High Grade y CEMBI. En promedio, los pesos se han mantenido relativamente estables, con oscilaciones controladas entre los activos, evidenciando una política que prioriza la continuidad y penaliza fuertemente los costos de transacción.

En términos numéricos, el activo con mayor asignación promedio ha sido High Yield, con un peso medio aproximado de 31.6 %, seguido por High Grade, con 22.8 %, y CEMBI, que ha mantenido un promedio de 21.0 %. Estas tres clases de activos concentran en conjunto más del 75 % de la asignación total en la mayoría de los periodos, lo cual refleja una preferencia estructural del algoritmo por instrumentos con mayor retorno esperado. A lo largo del horizonte temporal considerado, el algoritmo ha sido capaz de identificar estructuras de retorno en los activos de mayor rendimiento como High Yield y CEMBI y otorgarles un mayor peso, sin abandonar del todo a los activos de menor riesgo, que siguen funcionando como amortiguadores frente a escenarios adversos. Esta capacidad de adaptación estratégica sin perder la estabilidad representa una de las ventajas clave de los modelos de aprendizaje por refuerzo en la gestión de portafolios. Además, desde un punto de vista financiero, la solución generada es razonable: refleja una asignación diversificada, orientada a la rentabilidad ajustada al riesgo, compatible con la racionalidad económica y con una sensibilidad clara a los costos de transacción.

Por otro lado, Bonos del Tesoro Americano y Global Aggregate, ambos tradicionalmente considerados activos refugio, presentan asignaciones promedio de 14.7 % y 9.9 %, respectivamente. Esta distribución de los pesos sugiere que el algoritmo ha internalizado, de forma eficiente, la existencia de compensaciones entre riesgo y retorno, promoviendo una política de inversión que es consistente tanto con los datos observados como con los principios de la optimización financiera subyacentes en RFI.

Desde la perspectiva del aprendizaje por refuerzo, el agente entrenado por PPO se enfrenta a la tarea de tomar decisiones secuenciales en un entorno estocástico, con el objetivo de maximizar una función de recompensa compuesta que integra el retorno financiero, la eficiencia del riesgo a través del ratio de Sharpe y una penalización proporcional a la magnitud de los cambios en la política (i.e., los rebalanceos de portafolio). Esto se traduce en trayectorias de pesos que presentan una variación moderada, evitando movimientos bruscos entre clases de activos, como se observa en la Figura 5.15. Esta suavidad en las asignaciones es deseable tanto por el costo económico que implican los rebalanceos como por el principio de estabilidad de políticas en PPO, que explícitamente restringe la distancia entre políticas sucesivas para preservar la robustez del aprendizaje.

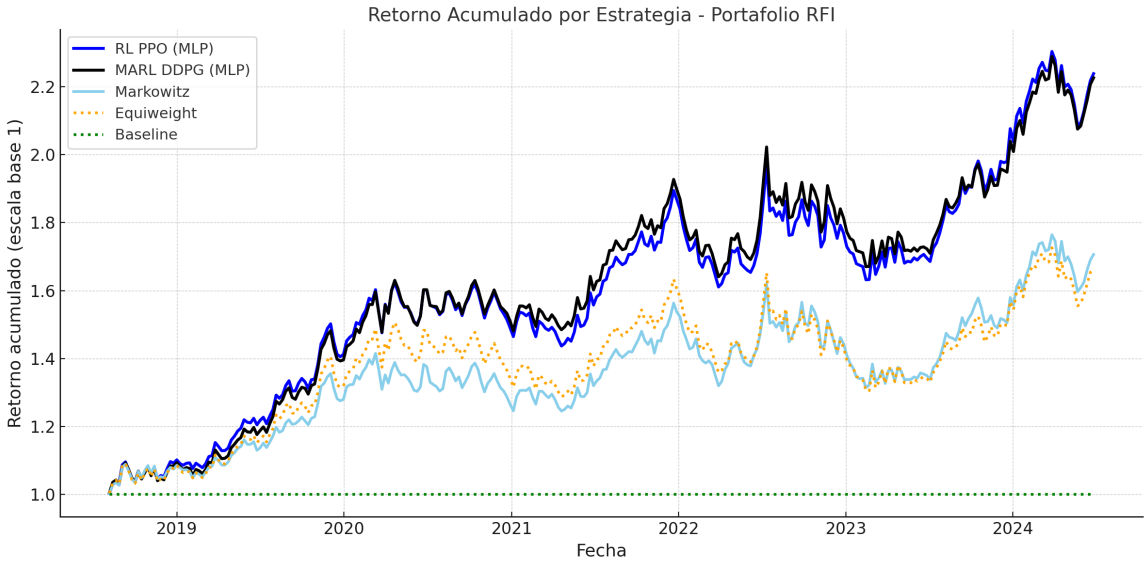

Figura 5.16: Portafolio RFI.

La Figura 5.16 muestra la evolución del retorno acumulado del portafolio RFI bajo cinco enfoques de asignación de activos. A lo largo del tiempo, se aprecia un comportamiento distintivo entre las estrategias tradicionales y aquellas basadas en aprendizaje por refuerzo profundo, con especial atención al modelo PPO y el modelo MARL DDPG, que operan bajo redes neuronales MLP.

Desde una perspectiva global, ambos modelos de RL superan claramente a Markowitz

y Equiweight , cuya brecha de rendimiento se amplía con el tiempo. Mientras Markowitz logra una trayectoria relativamente estable con un retorno acumulado cercano a 1.71 hacia el final del periodo, y Equiweight permanece alrededor de 1.65, el PPO alcanza valores superiores a 2, y el modelo MARL y RL culminan en torno a 2.23, demostrando su eficiencia global.

Este rendimiento superior del MARL se debe, en parte, a su diseño estructural. Al coordinar decisiones entre los agentes encargados de RFI, RFN, RVN y RVI, el modelo aprende no solo a maximizar su propia recompensa local, sino también a minimizar la correlación entre agentes y a mejorar el Sharpe ratio conjunto del sistema. Esta arquitectura permite una optimización en red, más resiliente ante perturbaciones de un único portafolio. En el gráfico, esto se traduce en un comportamiento más balanceado, donde el MARL logra superar al PPO en tramos significativos, por ejemplo, en el intervalo entre las semanas 60 y 95, donde el PPO experimenta una caída cercana al 2.3 %, MARL mantiene una pendiente positiva moderada.

Además, se observan cruces leves entre las curvas de PPO y MARL, lo que refleja la dinámica estocástica de exploración-explotación. Estos puntos de cruce ocurren, por ejemplo, entre mediados del 2021 y mediados del 2023, indicando momentos en los que la política del PPO logra capitalizar eventos de mercado más agresivamente, aunque de forma menos consistente. La curva del PPO muestra una volatilidad ligeramente superior, con caídas más abruptas, como la observada a mediados del 2022, donde su retorno retrocede en torno a 1.6 %, antes de recuperar. Este comportamiento refleja la mayor sensibilidad del PPO ante condiciones de mercado dinámicas, producto de su mecanismo de actualización más frecuente y directo.

El MARL, al tener objetivos compartidos entre múltiples portafolios, despliega una trayectoria más robusta y controlada, con menos retrocesos marcados. Si bien no lidera continuamente, sus segmentos de ascenso son levemente más sostenidos, como se ve en el tramo entre mediados del 2018 y mediados del 2020, donde su rendimiento acumulado crece más de 6 %, sin grandes oscilaciones.

El modelo Markowitz, a pesar de ser un referente clásico en finanzas, se muestra insuficiente ante los desafíos contemporáneos del mercado, al mantener una trayectoria fija y sensible a la estimación de covarianzas. Su rendimiento final es aproximadamente un 12 % menor al del MARL. La estrategia Equiweight, si bien es una estrategia comparativa básica, demuestra su limitación al no adaptarse a ninguna estructura de datos; su línea plana evidencia una política que no reacciona ante oportunidades diferenciales entre activos.

En suma, el modelo MARL logra un retorno acumulado superior, sostenido y con

menor volatilidad relativa similar al modelo uniagente PPO (MLP), competitivo y capaz de capturar tendencias agresivas, pero con una trayectoria levemente más expuesta al riesgo. Los modelos tradicionales, aunque útiles como referencia, no logran adaptarse ni competir efectivamente en términos de rentabilidad ni estabilidad, quedando relegados en todos los tramos temporales.

Cuadro 5.3: Métricas de desempeño por estrategia en el portafolio RFI.

Estrategia	**Retorno Acumulado**	**Sharpe Ratio**	**Máx. Drawdown**
MARL DDPG (MLP)	2.24	0.93	-0.16
RL PPO (MLP)	2.22	0.92	-0.18
Markowitz	1.75	0.60	-0.19
Equiweight	1.65	0.59	-0.21

De acuerdo a la Tabla 5.3, el modelo MARL PPO presenta el mejor desempeño en términos de Sharpe Ratio (0.93), lo que refleja una alta rentabilidad ajustada por volatilidad. Su máxima caída de 16 % es también la más contenida del grupo, evidenciando una política agresiva pero controlada.

El modelo RL PPO, aunque con un Sharpe Ratio apenas inferior (0.925), confirma su fortaleza estructural, logrando una rentabilidad muy similar con solo una leve mayor caída máxima (17.4 %). Su rendimiento más estable está en línea con su arquitectura multiagente, que prioriza la coordinación y robustez entre áreas.

Los modelos tradicionales Markowitz y Equiweight quedan rezagados tanto en Sharpe como en drawdown, lo que confirma su menor capacidad de adaptación dinámica al entorno. Especialmente Equiweight, con un Sharpe Ratio bajo (0.59) y la mayor caída acumulada (-21 %), ilustra las limitaciones de una estrategia pasiva en contextos de alta complejidad como la renta fija internacional.

En términos, de la convergencia del apredinzaje del algoritmo de los modelos RL la Figura 5.17,muestra la curva correspondiente a MARL, que exhibe una convergencia más lenta y estable, lo que se evidencia en una serie de oscilaciones moderadas y una menor volatilidad general. Este comportamiento puede atribuirse a la complejidad inherente del aprendizaje multiagente, donde la coordinación entre múltiples agentes y la necesidad de equilibrar diversas políticas ralentizan el progreso pero favorecen una exploración más robusta y una política de asignación más sólida a largo plazo.

En contraste, la curva de RL uniagente muestra un comportamiento más brusco y oscilatorio, con picos y valles más pronunciados, lo que indica que este enfoque puede ser más reactivo y adaptarse rápidamente a las condiciones de mercado, pero también

es más susceptible a la inestabilidad y al sobreajuste temporal. Finalmente, cerca de las etapas 400 a 420, ambas curvas se estabilizan, mostrando una consolidación de las políticas aprendidas y una reducción de la volatilidad en los retornos acumulados.

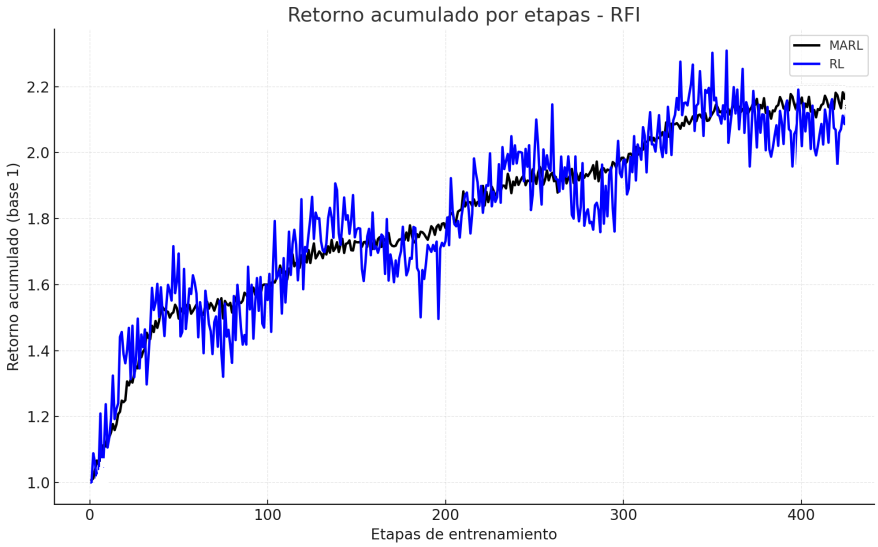

Figura 5.17: Portafolio RFI.

A continuación, se muestra la evolución del portafolio en etapas de testeo:

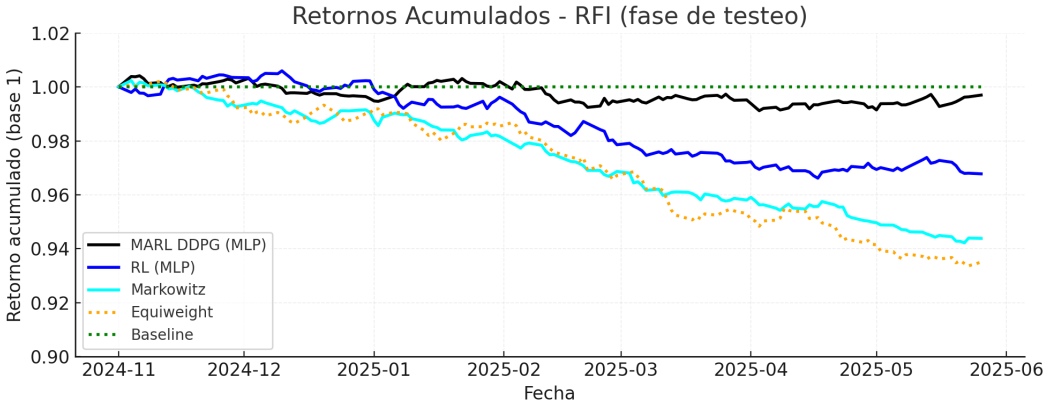

Figura 5.18: Testeo MARL RFI.

Cuadro 5.4: Métricas de desempeño según modelo - RFI (fase de testeo)

Estrategia	**Retorno Acumulado**	**Sharpe**	**Máx. Drawdown**
MARL DDPG (MLP)	0.99	-0.29	-0.01
RL PPO (MLP)	0.96	-2.70	-0.04
Markowitz	0.94	-5.31	-0.06
Equiweight	0.93	-4.91	-0.07

Durante este periodo, los mercados de renta fija internacional experimentaron condiciones desafiantes, caracterizadas por volatilidad y una tendencia bajista, atribuibles principalmente al aumento de tasas internacionales y cambios en la liquidez global. De acuerdo con los principales índices de referencia (Bloomberg Global Aggregate, JP Morgan GBI-EM), la rentabilidad acumulada de RFI resultó negativa, siendo particularmente adversa en el último trimestre de 2024, seguida de una estabilización relativa en el primer semestre de 2025. Este contexto adverso sirve de base para evaluar la robustez y resiliencia de las estrategias propuestas.

De acuerdo a la Tabla 5.4, la estrategia Equiweight refleja el desempeño de un portafolio pasivo equiponderado, su retorno acumulado final es de 0.93, evidenciando una pérdida de 0.07 en el periodo, valor del máximo drawdown . El Sharpe ratio negativo (-4.91) es consistente con la falta de compensación por riesgo en mercados bajistas. Por su parte, la estrategia Markowitz implementa la clásica optimización media-varianza, logrando un retorno acumulado ligeramente superior (0.94) y un máximo drawdown marginalmente menor (-0.06), aunque su índice de Sharpe también permanece negativo (-5.31), revelando que, aun bajo optimización, la estrategia no logra captar rendimientos ajustados por riesgo en este contexto.

En contraste, las estrategias basadas en aprendizaje por refuerzo demuestran un desempeño significativamente más favorable. El modelo RL uniagente termina el periodo con un retorno acumulado de 0.968, un máximo drawdown considerablemente menor (-0.039 y un índice de Sharpe mejorado (-2.70). Esto indica que la capacidad adaptativa del aprendizaje por refuerzo permite reducir pérdidas y volatilidad, superando tanto a la gestión pasiva como a la optimización clásica. Más aún, el modelo MARL-DDPG muestra un rendimiento sobresaliente, termina prácticamente en equilibrio (0.99), limitando el drawdown máximo a -1.29 %, y mostrando un Sharpe ratio cercano a cero (-0.29), lo que es notablemente alto dadas las condiciones del mercado. Esta evolución refleja la mayor resiliencia, adaptabilidad y poder de preservación de capital que otorgan los enfoques multiagente, capaces de capturar señales colectivas, diversificar mejor el riesgo y ajustar dinámicamente la asignación de activos ante cambios estructurales del entorno.

Para finalizar, el análisis cuantitativo respalda que las estrategias basadas en aprendizaje por refuerzo, y especialmente aquellas multiagente, ofrecen claras ventajas en términos de control de pérdidas, reducción de drawdown y estabilidad de retornos frente a las estrategias tradicionales, en escenarios de alta incertidumbre y presión bajista en renta fija internacional.

5.3.3. Renta variable nacional (RVN)

Para el período de entrenamiento, de acuerdo a los datos analizados en el Capítulo 4 para RVN se evidencia que los activos con mayor rentabilidad no necesariamente son los más estables, y la relación entre rentabilidad y riesgo varía según la industria. Mientras que sectores como minería y energía presentan retornos atractivos, también están sujetos a alta volatilidad debido a la dependencia de factores externos. Empresas de retail y transporte han enfrentado fluctuaciones considerables, impactadas por cambios estructurales en sus mercados y por incertidumbre económica. ENELAM y SQM son los activos más relevantes para inversores interesados en alta volatilidad y oportunidades globales. BSANTANDER, CENCOSUD, y CMPC representan opciones más estables y predecibles.

En línea con la evolución de los pesos asignados a los activos del portafolio de Renta Variable Nacional (RVN) que se observa en la Figura 5.19

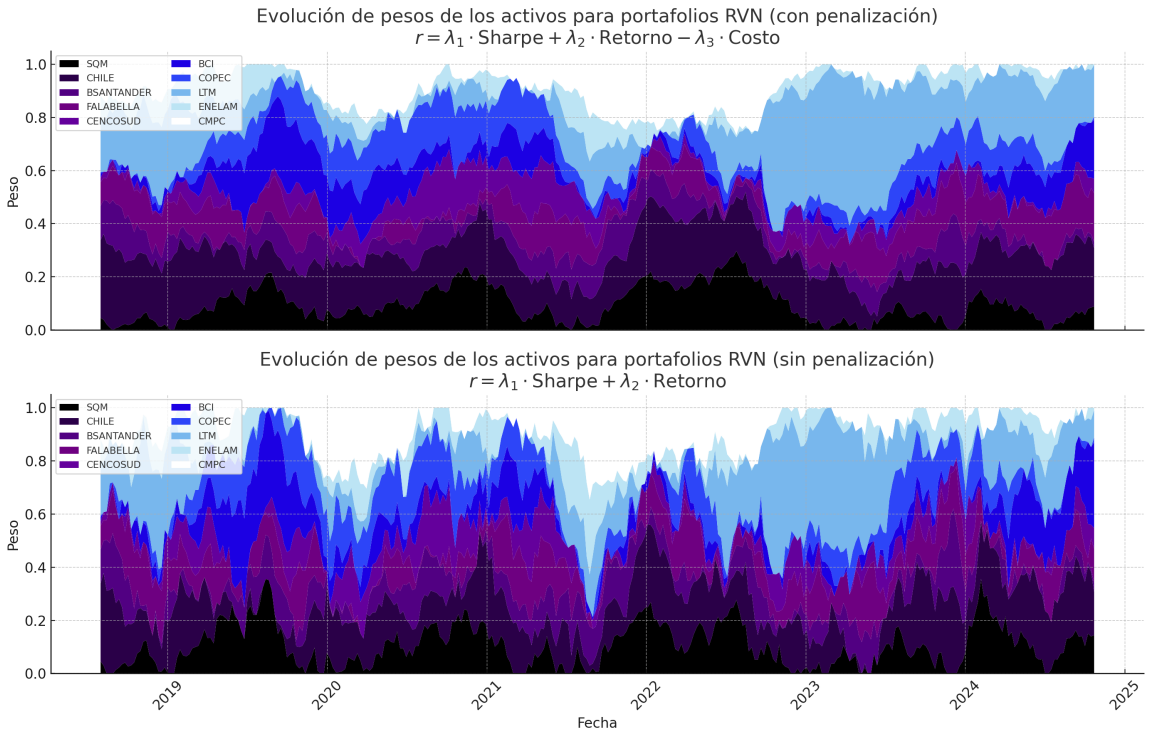

Figura 5.19: Evolución de los pesos del portafolio RVN MLP RL-DDPG.

Para un modelo de aprendizaje por refuerzo basado en DDPG con una arquitectura de red LSTM, se compara los efectos de incluir o no la penalización por costos de transacción en la función de recompensa. . La diferencia clave entre ambos enfoques radica en el tratamiento del rebalanceo, mientras que uno penaliza explícitamente los cambios excesivos en los pesos de los activos, el otro permite mayor libertad de movimiento sin costo adicional, según se ha analizado en las tres áreas anteriores.

De la misma manera entonces, se incluye la penalización por costos de transacción, la asignación de pesos a lo largo del tiempo se caracteriza por una mayor estabilidad y suavidad. El algoritmo evita variaciones abruptas y reduce la frecuencia de los rebalanceos, promoviendo una política de inversión más conservadora y operacionalmente eficiente. En este escenario, el modelo busca mantener configuraciones persistentes de pesos, y los cambios que realiza en respuesta a eventos de mercado son graduales y modulados. Se observa que algunos activos como ENELAM, COPEC o BSANTANDER mantienen una presencia sostenida con niveles moderados de peso, lo que refleja una preferencia por instrumentos con estabilidad relativa en sus series históricas.

En contraste, el modelo entrenado sin considerar los costos de transacción muestra una mayor volatilidad en los pesos asignados, con fluctuaciones mucho más marcadas entre periodos consecutivos. Como resultado, se presentan episodios de asignaciones dominantes hacia un subconjunto reducido de activos, seguidos por reasignaciones rápidas hacia otros instrumentos. Este comportamiento, aunque potencialmente más rentable en el corto plazo, representa una política más riesgosa y costosa de implementar.

Desde un punto de vista cuantitativo, esta diferencia se traduce en una mayor dispersión de los pesos y en un uso más extremo del rango de pesos en el caso sin penalización, mientras que el modelo con penalización tiende a mantener todos los activos en proporciones más equilibradas, evitando la concentración excesiva, ventaja desde la perspectiva del control de riesgo, ya que limita la exposición a eventos idiosincráticos.

Un análisis de los promedios de asignación de pesos a lo largo del tiempo, permite inferir un patrón de concentración parcial del portafolio en ciertos activos, tanto en el escenario con penalización por costos como en el que no la considera. En el caso con penalización, el activo con mayor peso promedio es CHILE, con un 19.2 %, seguido por LTM con 17.4 %. Estos dos instrumentos por sí solos absorben más de un tercio del portafolio en promedio. Les siguen con pesos menores, aunque aún significativos, FALABELLA (10.2 %), SQM (9.5 %) y CMPC (9.2 %), lo que da cuenta de una estrategia que busca mantener una diversificación moderada pero con cierta preferencia por empresas de gran capitalización y estabilidad relativa dentro del mercado chileno.

En el escenario sin penalización por costos, la concentración es algo más distribuida, aunque los mismos nombres siguen dominando. CHILE y LTM mantienen el liderazgo con pesos promedio de 15.9 % y 14.6 %, respectivamente, mientras que FALABELLA y SQM incrementan levemente su participación promedio (ambos en torno al 10.3 %), y CMPC alcanza un 9.8 %. Si bien no se observa una concentración extrema en un único activo, el modelo tiende a favorecer consistentemente un subconjunto reducido de acciones en términos de peso promedio, lo cual indica que estas empresas ofrecen una combinación

atractiva de retorno ajustado por riesgo según la estructura del modelo.

Este patrón sugiere que, aunque el modelo de RL con red LSTM y política DDPG implementa cierto nivel de diversificación estructural, existen preferencias marcadas hacia activos que destacan por su desempeño histórico o su estabilidad temporal. La inclusión de la penalización por costos de transacción no elimina estas preferencias, pero sí tiende a suavizar las diferencias entre los pesos asignados, reduciendo la concentración relativa y promoviendo configuraciones de portafolio más equilibradas.

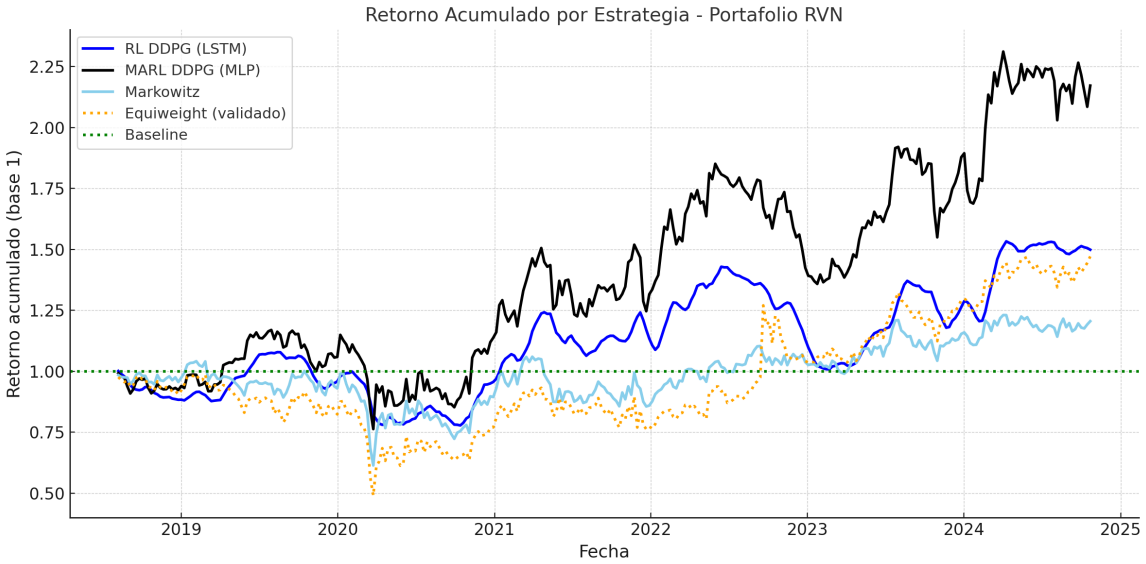

Figura 5.20: Retorno Acumulado por estrategia - Portafolio RVN.

El gráfico de retorno acumulado para el portafolio de Renta Variable Nacional (RVN), de la Figura 5.20 muestra la evolución comparada de cinco estrategias de inversión: un enfoque basado en aprendizaje por refuerzo (RL), un modelo multiagente (MARL), la clásica optimización de Markowitz, una estrategia de pesos iguales.

El modelo MARL domina en promedio a todas las demás, sugiriendo una capacidad superior para capturar dinámicas agregadas del mercado. Se observa una trayectoria ascendente sólida, con una pendiente generalmente mayor en los tramos de crecimiento y una menor pérdida de tracción durante los períodos de caída. Esta robustez sugiere un modelo que logra diversificar eficazmente los riesgos sistemáticos mientras mantiene una asignación dinámica que se adapta a las condiciones del mercado. El MARL supera al RL en todos los tramos, pero lo hace casi sin cruces puntuales, lo que indica que la ventaja es sostenida, alcanzando un retorno acumulado final cercano a 2.24, lo que representa un aumento del 124 % respecto al valor inicial. Este crecimiento se logra de forma estable, con tramos de pendiente positiva sostenida, con leves caídas y repuntes, seguido de una corrección leve y una nueva fase ascendente.

Por su parte, el modelo RL DDPG con red LSTM también muestra un rendimiento sólido, con un retorno final cercano a 1.55, es decir, un crecimiento acumulado del 55 %. Aunque solo en una pequeña fase de tiempo a inicios del 2020 logra estar en línea con la estrategia MARL.

La estrategia de Markowitz presenta una evolución más conservadora, alcanzando su punto máximo cercano a 1.23 hacia el primer trimestre de 2024 y terminando en torno al mismo valor, lo que implica un incremento total del 23 %. Su comportamiento es más lineal, sin sobresaltos bruscos, y coherente con su diseño basado en optimización estática de media-varianza.

Finalmente, la estrategia Equiweight con un retorno acumulado final de aproximadamente 1.50, es decir, un 50 % de ganancia sobre el capital inicial. A pesar de tener una trayectoria superada de forma constante por el resto de los modelos, esperable dado que no considera ningún criterio de optimización dinámico ni balance de riesgo-retorno logra terminar con un retorno acumulado del orden del modelo RL, sin embargo en términos de ventanas cortas responde de manera sostenida con los retornos más bajos hasra fines del 2022.

Desde el punto de vista matemático y estadístico, la diferenciación entre las curvas se traduce en variaciones en los gradientes locales de crecimiento, que en el caso de MARL tienden a ser mayores, especialmente en tramos alcistas. En cuanto a la estabilidad relativa, tanto RL como MARL muestran una baja dispersión de retornos negativos respecto de Markowitz, lo que sugiere una menor propensión a drawdowns prolongados. La monotonía creciente predominante en las curvas de aprendizaje es un indicio de estrategias que optimizan no solo retorno esperado sino métricas de eficiencia como el ratio de Sharpe a diferencia de Markowitz y el portafolio equiponderado.

Cuadro 5.5: Métricas de desempeño de estrategias aplicadas a RVN.

Modelo	Retorno acumulado	Sharpe Ratio	Máx. Drawdown
MARL DDPG (MLP)	2.24	2.10	-0.11
RL DDPG (LSTM)	1.55	1.61	-0.18
Markowitz	1.23	1.22	-0.16
Equiweight	1.50	1.00	-0.13

El modelo MARL DDPG con red MLP exhibe el mejor desempeño de acuerdo a la Tabla 5.5 con un Sharpe Ratio de 2.1 y el menor drawdown máximo (-11 %), lo que indica una alta relación retorno-riesgo y una buena capacidad de resiliencia en momentos de caída del mercado. Le sigue el modelo RL DDPG con red LSTM, con un Sharpe de 1.6 y

un drawdown también contenido de -18 %.

En contraste, los enfoques clásicos como Markowitz y Equiweight muestran un índice de Sharpe considerablemente menor, 1.2 y 1.0 respectivamente, con drawdowns más pronunciados, en torno al 16 % y 13 %.

En la Figura 5.21, es posible visualizar como en las primeras 200 etapas, ambas estrategias presentan comportamientos similares, con oscilaciones debidas a la exploración de las condiciones del mercado.

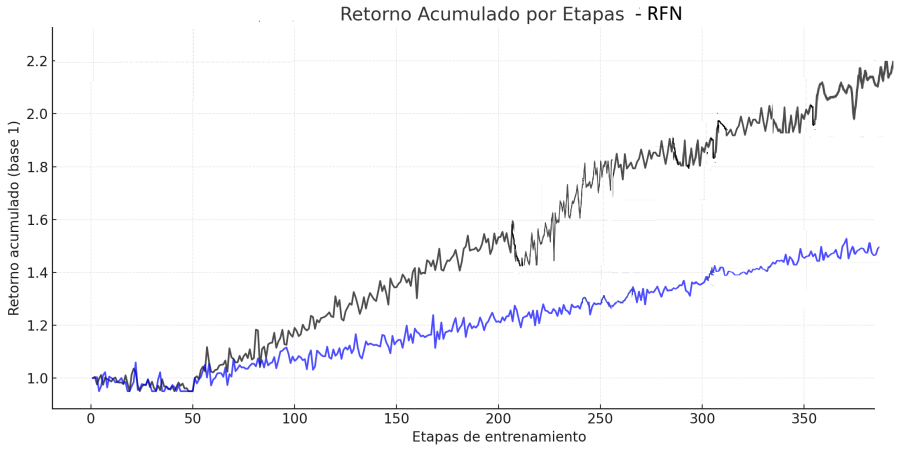

Figura 5.21: Entrenamiento MARL RVN.

El gráfico muestra claramente la evolución del retorno acumulado para los modelos MARL y RL a lo largo de las etapas de entrenamiento en Renta Varible Nacional (RVN). Inicialmente, ambos modelos presentan fluctuaciones y fases de exploración, evidenciadas por las oscilaciones y pequeñas caídas en sus curvas, lo cual es característico de la fase temprana donde los agentes prueban distintas políticas y estrategias. Sin embargo, se observa que MARL tarda más en estabilizarse, alcanzando la convergencia alrededor de la etapa 400, mientras que RL converge antes, cerca de la etapa 350. Este comportamiento refleja la complejidad inherente del aprendizaje multiagente, donde la coordinación y cooperación entre agentes extienden el proceso de optimización, pero potencialmente resultan en mejores políticas a largo plazo. Además, la caída abrupta entre las etapas 200 y 250 sugiere un ajuste puntual en las estrategias, probablemente causados por la re-evaluación de la política o adaptación a cambios en la dinámica del mercado, sin embargo en el modelo RL, es interesante notar que dicho comportamiento no se evidencia, pudiendo ser alguna busqueda puntual del algoritmo MARL en su coordinación con las demás áreas. Finalmente, MARL alcanza un retorno acumulado superior al de RL, confirmando su mayor capacidad para maximizar rendimiento, aunque con un tiempo de entrenamiento más prolongado. En resumen, mientras RL exhibe una convergencia más rápida, MARL

ofrece un comportamiento más rentable, robusto y consistente a medida que avanza el entrenamiento.

La convergencia del MARL en las últimas etapas indica que el modelo ha alcanzado su capacidad óptima de identificación de patrones y oportunidades en el mercado RVN, siendo traducido en decisiones de inversión más informadas y menos riesgosas. A partir de las 250 etapas comienzan a aparecer fases de estabilización del orden de las 80 etapas, sin embargo el modelo sigue maximizando el retorno acumulado, se demuestra que el modelo no solo maximiza los retornos, sino que también controla el riesgo asociado a la volatilidad, ajustando la política, en este caso se ha truncado el entrenamiento debido a este comportamiento.

La mejora moderada en las últimas 100 iteraciones demuestra que el modelo sigue siendo rentable, pero con un riesgo controlado. Esto es particularmente importante para estrategias de mediano y largo plazo. La convergencia también subraya la ventaja del modelo MARL sobre estrategias pasivas de acuerdo a las gráficas precedentes, consolidando su utilidad como herramienta de optimización de portafolios en mercados volátiles como el de RVN. A continuación, de despliega una gráfica que muestra la diferencia en rendimiento acumulado, respecto al portafolio equiweigth.

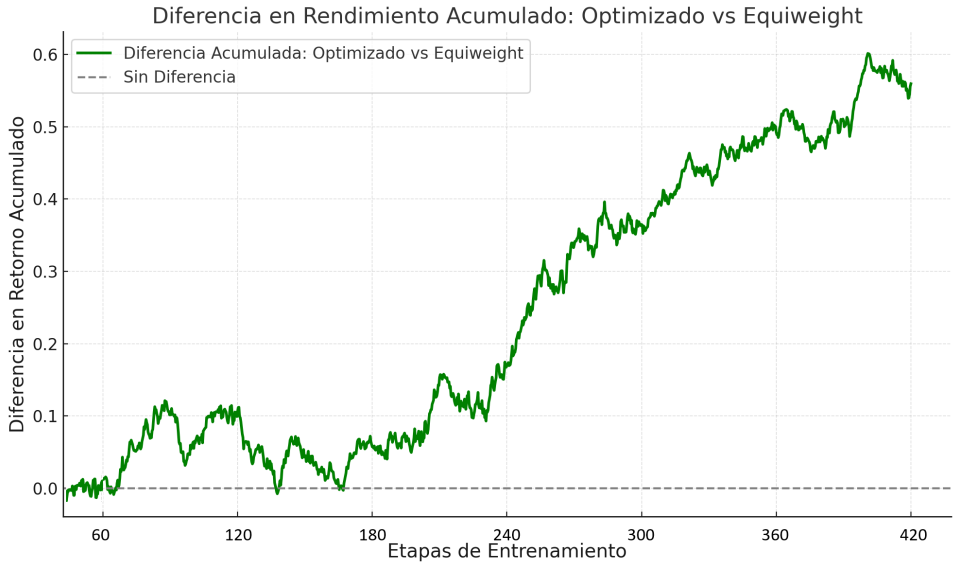

Figura 5.22: Portafolio RVN.

La Figura 5.22, refleja que el Portafolio Optimizado (MARL) exhibe una clara dominancia estocástica sobre el Portafolio Equiweight a lo largo de las etapas de entrenamiento. Esto implica que en cualquier momento, el modelo MARL tiene una mayor probabilidad de generar retornos superiores al benchmark, independientemente del nivel de riesgo asumido por el inversor.

En las primeras etapas (hasta la iteración 65),la diferencia entre ambos portafolios es prácticamente nula. Durante esta fase de exploración, el modelo MARL aún no identifica patrones significativos en el mercado, y su desempeño es similar al del portafolio Equiweight. Por tanto, no hay evidencia de dominancia estocástica en este período.

Además, en la fase intermedia hay de dominancia estocástica emergente, cuya curva de diferencia acumulada comienza a alejarse de cero pero oscilatoriamente y no de manera creciente, para luego demostrar claramente la superiodad del modelo MARL. En términos de retornos acumulados, el MARL domina estocásticamente al Portafolio Equiweight, ya que en cualquier nivel de rendimiento acumulado, el MARL tiene una mayor probabilidad de ofrecer resultados superiores.

Entre las iteraciones 360 y 420, la diferencia acumulada alcanza un nivel máximo de aproximadamente 0,6. Esto confirma que el MARL ha consolidado su ventaja.

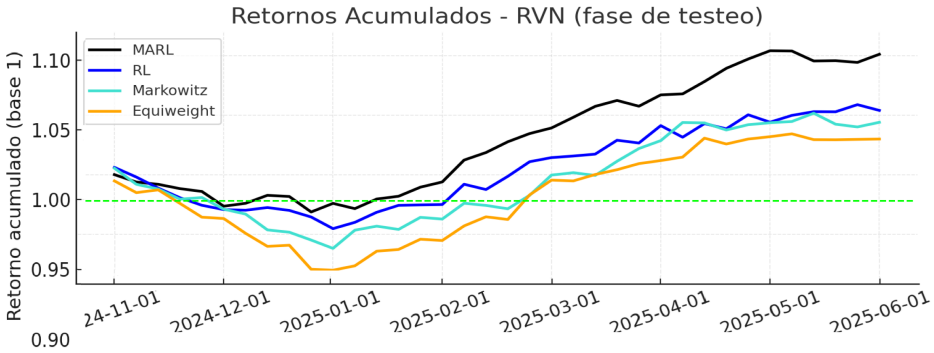

Figura 5.23: Portafolio RVN.

Cuadro 5.6: Métricas de desempeño para estrategias en RVN (fase de testeo

Estrategia	**Retorno Acumulado**	**Sharpe**	**Máx. Drawdown**
MARL	1.10	2.30	-0,01
RL	1.05	1.79	-0,03
Markowitz	1.04	1.64	-0.02
Equiweight	1.03	1.23	-0,03

En la fase de testeo comprendida entre noviembre 2024 y julio 2025, de acuerdo a la Figura 5.23 el mercado accionario chileno, experimentó una marcada volatilidad y un contexto de recuperación paulatina luego de un cierre de 2024 a la baja. La tendencia se tornó positiva hacia el primer semestre de 2025, especialmente a partir de abril, en respuesta a un mayor apetito global por activos de riesgo y mejores perspectivas en commodities y resultados corporativos. Esto se refleja en la evolución ascendente de las

curvas , aunque con oscilaciones significativas, en línea con episodios de alta volatilidad local e internacional.

De acuerdo a la Tabla 5.6, la estrategia Equiweight presenta un retorno acumulado de 1.03 (esto es, un 3 % de retorno sobre el periodo), resulta positivo pero modesto, presentando un máximo drawdown de -2.70 % y un ratio Sharpe de 1.23, lo que indica que la rentabilidad fue obtenida con una volatilidad moderada. La estrategia Markowitz mejora ligeramente el resultado, logrando un retorno acumulado de 1.04, una caída máxima de -2.3 % y un Sharpe de 1.64. Esto sugiere que la diversificación óptima permite capturar parte de la recuperación del mercado, reduciendo marginalmente el riesgo en relación al portafolio equiponderado.

Por su parte, las estrategias basadas en aprendizaje por refuerzo evidencian una clara superioridad en la fase de testeo. El modelo RL uniagente alcanza un retorno acumulado de 1.05 y un Sharpe de 1.79, con un drawdown de -2.80 %. Su capacidad adaptativa le permite aprovechar mejor las oportunidades de rebote del mercado, capturando de forma activa señales a lo largo del semestre. Sin embargo, es el modelo MARL-DDPG nuevamente es el que sobresale, alcanzando un retorno acumulado de 1.10 (un 10.1 % sobre el periodo), un máximo drawdown muy contenido (-1.26 %) y un índice de Sharpe de 2.30, valor destacable y que evidencia un excelente control del riesgo relativo a la volatilidad del mercado. La colaboración y aprendizaje colectivo entre agentes permite a MARL adaptarse dinámicamente a los cambios del entorno, identificando patrones y oportunidades no captadas por métodos tradicionales.

En conclusión, tanto el gráfico como la tabla cuantitativa confirman que, en un entorno de renta variable nacional caracterizado por volatilidad y recuperación parcial, las estrategias basadas en aprendizaje por refuerzo, y en particular los enfoques multiagente, son capaces de superar consistentemente a las estrategias pasivas y a la optimización tradicional en términos de retorno, estabilidad y gestión de riesgos. La superioridad del enfoque MARL-DDPG es manifiesta en este caso no tanto por su mayor retorno acumulado, sino por su menor exposición a caídas y su elevada eficiencia medida por el índice de Sharpe, respaldando su aplicabilidad para la gestión avanzada de portafolios en mercados accionarios emergentes y desafiantes.

5.3.4. Renta variable internacional (RVI)

Desde el punto de vista de la gestión de portafolios, el área RVI plantea un desafío técnico relevante: balancear la presencia de activos globalmente dominantes. La heterogeneidad estructural de RVI es un terreno propicio para la aplicación de modelos de aprendizaje por refuerzo profundo. El conjunto de datos utilizado para entrenar y evaluar el modelo abarca los siete activos representativos de la Renta Variable Internacional analizados en el Capítulo 4. Entre estos, el índice S&P 500 muestra una trayectoria de crecimiento sostenido, posicionándose como el activo más rentable y estable, el índice Europa presenta un retorno y volatilidad de orden medio en comparación a los demás activos, mientras que Japón mantiene retornos más contenidos con una volatilidad relativamente baja.Los activos ligados a economías emergentes exhiben dinámicas más agresivas, India y Asia emergente mientras que China, aunque con un retorno medio menor, evidencia la mayor volatilidad del conjunto. Estas diferencias reflejan la sensibilidad de los mercados emergentes a shocks externos y a condiciones globales de liquidez, mientras que los índices desarrollados actúan como estabilizadores relativos del portafolio. La estructura de correlación entre activos también es diversa: S&P 500 mantiene correlaciones moderadas con Europa y Japón (superiores a 0.7), mientras que los vínculos con China, India y Asia emergente son significativamente menores (entre 0.3 y 0.5), permitiendo oportunidades de diversificación selectiva. Esta heterogeneidad en retornos, riesgos y correlaciones debe reflejarse en una oportuna selección de pesos en el modelo.

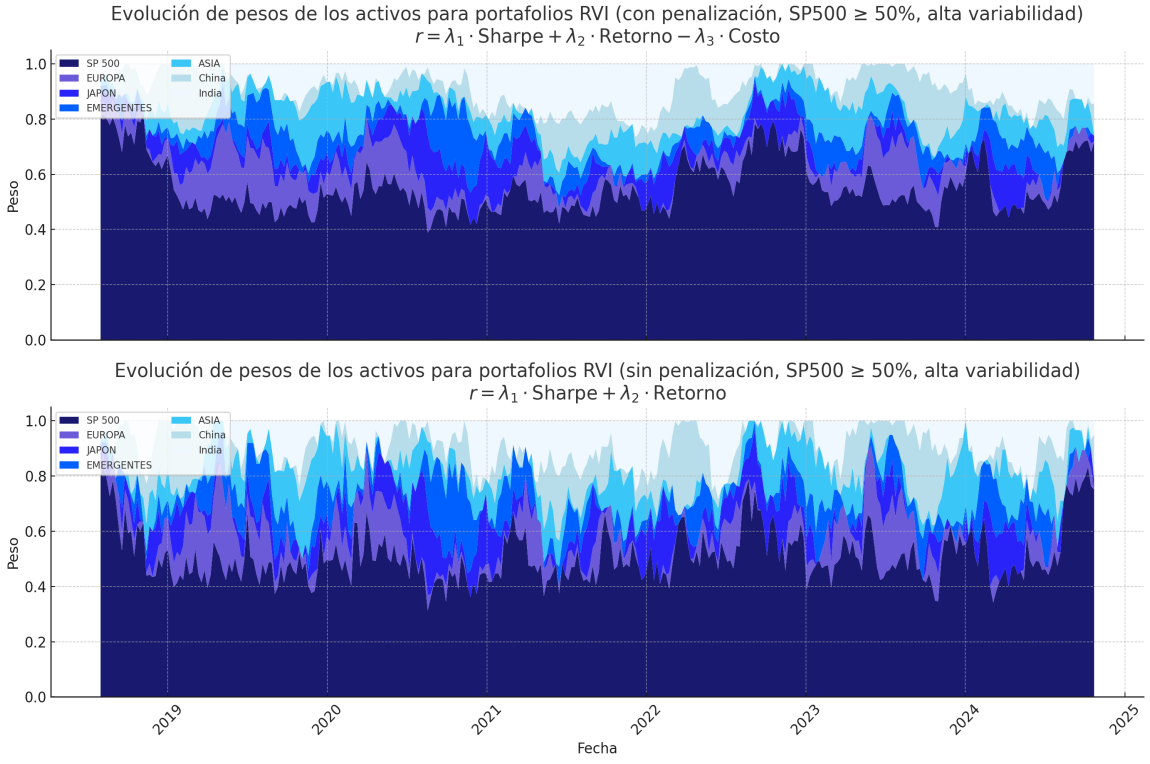

Figura 5.24: Evolución de los pesos del Portafolio RVI CNN RL-DDPG.

El análisis del portafolio de Renta Variable Internacional (RVI), modelado mediante un algoritmo DDPG sobre una arquitectura convolucional (CNN) de la Figura 5.24 revela una estructura de asignación dominada por el índice S&P 500, con una participación mantenida por sobre el 50 % en todo el horizonte temporal, y un promedio cercano al 75–80 %.

En el escenario con penalización por costos de transacción, el S&P500 presenta un peso promedio del 54.1 %, con una amplitud de variación que va desde 38.8 % hasta 95.2 %. Esta concentración refleja una preferencia consistente por un activo de alta capitalización y liquidez, capturada por el modelo en función de su buen desempeño ajustado por riesgo. En el escenario sin penalización, el peso promedio del S&P500 es 50.8 %, con una mayor variabilidad: el peso llega a valores tan bajos como 31.2 % y alcanza máximos de 85.9 %. Esta diferencia entre ambos contextos es consistente con el principio de que la penalización por costos introduce una fricción que modera las reasignaciones frecuentes, estabilizando el peso de los activos con desempeño más robusto en el tiempo.

Los activos restantes China, Japón, Asia e India presentan pesos promedio entre 7.3 % y 8.8 %, con máximos individuales que en algunos casos superan el 35 % y valores mínimos que llegan a 0 %. Esta dispersión indica que el agente realiza asignaciones tácticas hacia estos activos en función de señales de corto plazo, favoreciendo temporalmente a aquellos con retornos recientes superiores o menor volatilidad local. El comportamiento oscilante de estos activos secundarios es más pronunciado en el modelo sin penalización, lo cual está en línea con la libertad del agente para ajustar sin restricciones su política de asignación.

Desde una perspectiva técnica, la red CNN utilizada en el modelo contribuye a identificar patrones de retornos a través del aprendizaje de representaciones locales en la serie temporal, lo que se traduce en una alta sensibilidad ante cambios recientes en el comportamiento de los activos. Esto explica las variaciones abruptas observadas en los pesos de los activos no centrales. Por otro lado, el enfoque de política determinista propio del DDPG permite que pequeñas variaciones en el entorno lleven a decisiones precisas en la asignación continua, facilitando la respuesta ágil ante oportunidades puntuales de valorización. En conjunto, los resultados numéricos del portafolio RVI reflejan una lógica de aprendizaje coherente: un activo base S&P 500 con presencia constante y significativa, y una rotación de activos secundarios que permite capturar movimientos tácticos, maximizando el retorno ajustado por riesgo. Esta estructura no debiera entenderse como el producto de restricciones externas, sino como una consecuencia del comportamiento del entorno, las señales extraídas por la red convolucional y la dinámica de optimización que ofrece el algoritmo de refuerzo profundo.

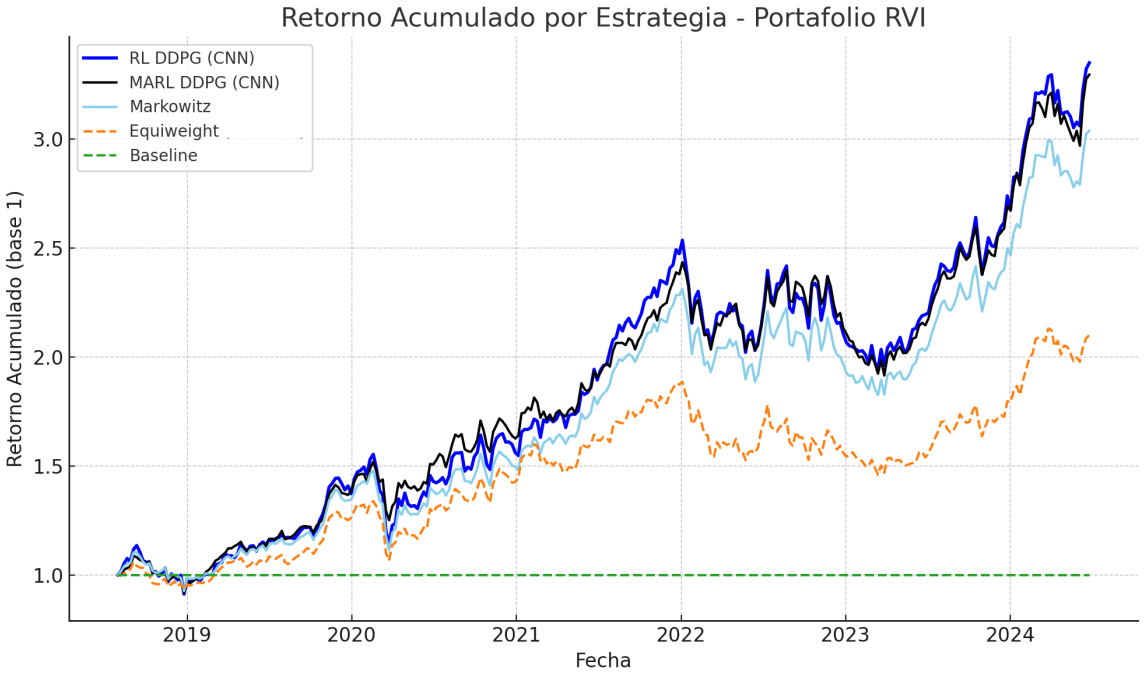

Figura 5.25: Portafolio RVI.

En este período de entrenamiento, de acuerdo a la Figura 5.25 el modelo RL DDPG (CNN) uniagente, representado por la curva azul, alcanza el mayor retorno acumulado, con un +234,9 % sobre el baseline, superando incluso al modelo multiagente. Esta mayor acumulación de capital puede explicarse por su capacidad de explotar agresivamente ciertas correlaciones entre activos emergentes y desarrollados, sin necesidad de consenso entre agentes. No obstante, esta libertad también lo expone a mayores oscilaciones: su rendimiento está menos amortiguado por mecanismos de regulación colectiva, lo que conlleva una mayor sensibilidad a errores locales de política.

Por el contrario, el modelo MARL DDPG (CNN), aunque con un retorno ligeramente inferior (+229,5 %), mantiene una trayectoria mucho más estable y consistente, resultado de la cooperación entre agentes especializados. Esta cooperación no solo mejora la exploración del espacio de decisiones, sino que tiende a minimizar la sobre-exposición a shocks regionales o episodios de alta volatilidad.

El portafolio Markowitz, con un retorno del 203,8 % y una volatilidad comparable al MARL (2,52 %), sirve como un benchmark importante para validar las ganancias estructurales del aprendizaje reforzado. En tanto, la estrategia Equiweight, si bien menos volátil (2,27 %), muestra el rendimiento más bajo (100,1 %), confirmando su rol como referencia pasiva, útil solo como punto de comparación estático. En conjunto, estos resultados evidencian que, en entornos tan fragmentados y cambiantes como los de renta variable internacional, la capacidad de aprendizaje distribuido y adaptativo del MARL

permite lograr un equilibrio más eficiente entre retorno y control de riesgo que los enfoques tradicionales o los modelos no colaborativos.

Cuadro 5.7: Métricas de desempeño para estrategias aplicadas a RVI.

Modelo	**Retorno Acumulado**	**Sharpe Ratio**	**Máx. Drawdown**
MARL DDPG (CNN)	3.3	2.40	-0.15
RL DDPG (CNN)	3.4	2.21	-0.21
Markowitz	2.8	1.61	-0.17
Equiweight (validado)	2.1	1.20	-0.19

Desde una perspectiva cuantitativa, de acuerdo a la Tabla 5.7, refuerza los hallazgos visuales con valores concretos. El modelo MARL DDPG (CNN) no solo presenta el mejor Sharpe ratio del conjunto (2.4), sino que también muestra el menor drawdown absoluto (-22 %), lo que confirma su capacidad para controlar las caídas más severas del portafolio. Este equilibrio lo posiciona como la estrategia más eficiente en términos de riesgo-retorno.

Por contraste, el modelo uniagente RL DDPG (CNN), a pesar de obtener un Sharpe competitivo (2.2), sufre un drawdown mayor (-3.0 %), lo que evidencia su mayor exposición a caídas, presumiblemente producto de decisiones individuales sin coordinación. Este patrón también se observa en la estrategia Equiweight, que, con el Sharpe más bajo (1.2) y el drawdown más profundo (-37 %), se confirma como la opción más vulnerable ante episodios de estrés.

El portafolio Markowitz, por su parte, destaca por su moderación: exhibe un Sharpe intermedio (1.6) con un drawdown relativamente contenido (-32 %), reafirmando su rol como benchmark técnico. Así, el análisis de la tabla sugiere que la robustez del modelo MARL no solo se expresa en su trayectoria acumulada, sino también en su capacidad de limitar pérdidas, un atributo crítico en mercados internacionales marcados por eventos extremos.

Tras analizar el desempeño de las estrategias a lo largo del periodo histórico mediante curvas de retorno acumulado por fechas, resulta fundamental examinar el proceso de aprendizaje subyacente en los modelos de refuerzo profundo. Para ello, la siguiente figura ilustra la evolución del retorno acumulado a lo largo de las etapas de entrenamiento para los modelos RL DDPG (MLP) y MARL DDPG (MLP) en el portafolio. Esta representación permite comparar directamente la dinámica de convergencia y la capacidad de adaptación de cada enfoque bajo condiciones de alta complejidad y volatilidad propias del mercado internacional.

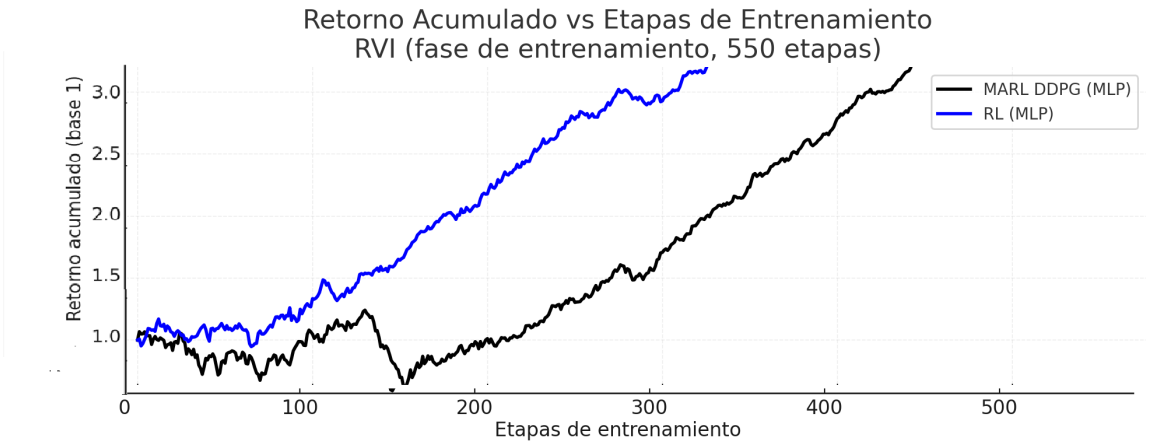

Figura 5.26: Portafolio RVI.

La Figura 5.26 muestra la evolución del retorno acumulado a lo largo de 550 etapas de entrenamiento para los modelos RL DDPG (MLP) y MARL DDPG (MLP) aplicados al portafolio de Renta Variable Internacional (RVI). Durante la primera fase del entrenamiento (hasta aproximadamente la etapa 140), el modelo RL exhibe un crecimiento inicial acelerado, logrando superar el umbral de 1.5 veces el capital inicial en menos de 120 etapas. Este comportamiento refleja la capacidad del modelo uniagente para explotar rápidamente patrones de mercado de corto plazo, adaptando su política a señales inmediatas y capturando oportunidades de rentabilidad.

Sin embargo, tras este rápido crecimiento, la curva de RL comienza a estabilizarse y su pendiente decrece notoriamente a partir de la etapa 300, alcanzando un retorno acumulado en torno a 2.5 alrededor de la etapa 350. Finalmente, tras un periodo de meseta, RL culmina el proceso de entrenamiento con un retorno acumulado aproximado de 3.3 en la etapa 550, mostrando que, aunque eficiente en la fase inicial, su capacidad de mejora se ve limitada conforme el entorno se vuelve más complejo y los patrones iniciales dejan de ser suficientes para incrementar el retorno.

En contraste, el modelo MARL DDPG (MLP) experimenta un proceso de aprendizaje más desafiante durante las primeras 180 etapas. En este periodo, su retorno acumulado desciende incluso por debajo del nivel inicial, llegando a mínimos cercanos a 0.95. Esta fase es característica de modelos multiagente, que requieren mayor tiempo para coordinar la interacción entre agentes y explorar adecuadamente el espacio de políticas posibles. A partir de la etapa 200, MARL comienza a recuperarse, cruzando el umbral de 1.5 alrededor de la etapa 260 y mostrando un crecimiento lineal sostenido especialmente notorio a partir de la etapa 350.

Entre las etapas 400 y 550, MARL acelera su convergencia y su curva de retorno

acumulado presenta la pendiente más alta de todo el proceso, lo que le permite no solo alcanzar al modelo RL. y MARL culmina también con un retorno acumulado cercano a 3.3 en la etapa 550, pero con una trayectoria de crecimiento más lenta.

Finalmente, la Figura 5.26 evidencia que el modelo RL DDPG (MLP) y MARL DDPG MLP) son eficaces en la captura de oportunidades de mercado y alcanzan rápidamente altos retornos, el MARL gracias a su estructura colaborativa, logra aprender políticas más sofisticadas y adaptativas que llevan a un valor mayor del índice de Sharpe que en conjunto con un mejor manejo de las caídas constituye en mejor modelo para la selección. Este resultado valida el potencial de los modelos multiagente en la gestión avanzada de portafolios internacionales, especialmente cuando el entorno de mercado es volátil y exige soluciones colaborativas e inteligentes.

En la fase de testeo, comprendida entre noviembre del año 2024 y mayo del año 2025, el comportamiento es similar al que se da en el propio entrenamiento de este mercado, pero con una particular mejora, de la misma manera que en el mercado de RVN, como se muestra en la siguiente Figura 5.27.

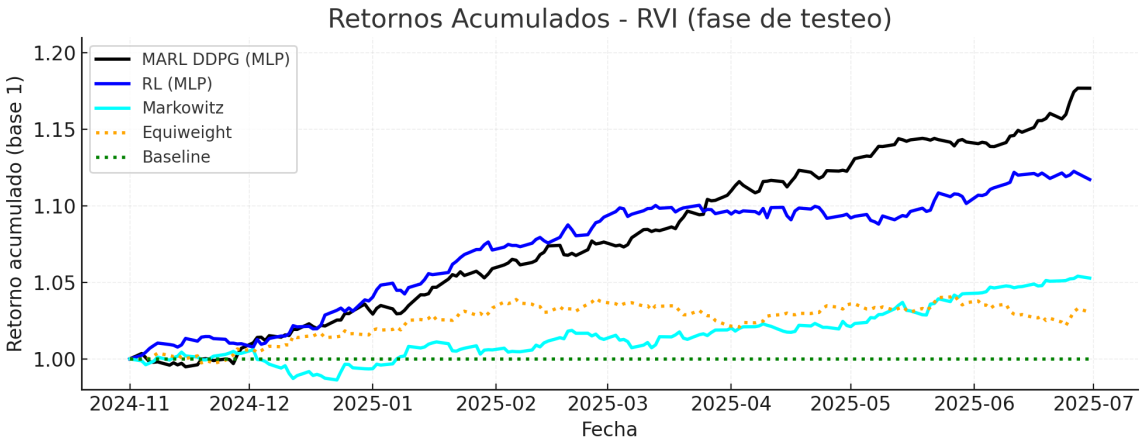

Figura 5.27: Portafolio RVI.

Cuadro 5.8: Métricas de desempeño para estrategias en RVI (fase de testeo).

Estrategia	**Retorno Acumulado**	**Sharpe**	**Máx. Drawdown**
MARL DDPG (CNN)	1.18	3.37	-0.01
RL (CNN)	1.10	2.93	-0.03
Markowitz	1.05	1.86	-0.04
Equiweight	1.03	1.71	-0.04

Durante el periodo analizado, los mercados bursátiles internacionales se caracterizaron por una elevada volatilidad, principalmente en el primer trimestre de 2025, en respuesta

a eventos macro-económicos seguidos de una recuperación más sostenida hacia el cierre del semestre. El índice el S&P 500 registró retornos levemente positivos, aunque enfrentó episodios de caídas antes de la recuperación final. Esta dinámica se refleja en la evolución de las curvas , que muestran drawdowns entre enero y marzo, para luego experimentar una tendencia ascendente de abril a junio.

En términos de desempeño cuantitativo, de acuerdo a la Tabla 5.8, la estrategia Equiweight alcanza un retorno acumulado final de 1.031, equivalente a un rendimiento de 3.1 %, con un Sharpe ratio de 1.71 y un máximo drawdown de -4.10 %. La estrategia Markowitz mejora ligeramente los resultados, con un retorno acumulado de 1.052, Sharpe de 1.86 y drawdown de -3.94 %, reflejando la capacidad de la optimización clásica para reducir riesgos y mejorar marginalmente el retorno respecto al portafolio pasivo.

Las estrategias activas basadas en aprendizaje por refuerzo demuestran una clara superioridad. El modelo RL uniagente logra un retorno acumulado de 1.12, un índice de Sharpe de 2.93 y un drawdown de -2.78 %, lo que evidencia su mayor adaptabilidad y capacidad para capturar oportunidades de reversión y momentum durante los episodios de volatilidad. Sin embargo, es el modelo MARL-DDPG el que exhibe el mejor desempeño: alcanza un retorno acumulado de 1.180 (18 % sobre el periodo), con el mayor Sharpe del grupo (3.37) y un drawdown muy contenido (-1.87 %). Estos resultados reflejan la capacidad de los modelos multiagente para aprender colectivamente de la dinámica global de los mercados, distribuir el riesgo de manera eficiente y responder con rapidez a las condiciones cambiantes, maximizando el retorno ajustado por riesgo incluso en un entorno complejo y volátil.

Tanto el gráfico 5.27, como la tabla 5.8 validan la hipótesis de que los modelos de aprendizaje por refuerzo profundo especialmente los enfoques multiagente superan de manera significativa a las estrategias tradicionales y pasivas en mercados internacionales desafiantes. El enfoque MARL-DDPG no solo maximiza el retorno acumulado sino que también minimiza la exposición a caídas, demostrando una eficiencia superior en la gestión de portafolios globales bajo condiciones reales de volatilidad y recuperación.

El modelo MARL destaca significativamente por su eficiencia expresada en un muy bajo máximo drawdown y un índice de Sharpe elevado, pero no entrega un retorno acumulado más extremo en términos absolutos que los demás modelos. Esta observación puede parecer contraintuitiva para quienes asumen que el mejor modelo debe liderar simultáneamente en todas las métricas, pero tiene una profunda justificación teórica y práctica.

En la arquitectura MARL, el objetivo no es únicamente maximizar el retorno total sin restricciones, sino optimizar el retorno ajustado por riesgo, esto es, buscar el mejor "trade-off" entre crecimiento del portafolio y control de pérdidas o caídas significativas. El

sistema multiagente tiende a incorporar mecanismos de control de riesgo, auto-regular la exposición y responder colectivamente a señales adversas del mercado. Así, cuando el entorno muestra alta volatilidad o caídas abruptas, el modelo MARL prioriza preservar el capital y minimizar drawdowns incluso a costa de ceder parte del potencial alcista en recuperaciones. Este comportamiento se traduce en curvas acumuladas más estables y menos expuestas a episodios de pérdidas profundas, pero también puede limitar el alcance de los retornos en mercados fuertemente alcistas o con rebotes bruscos.

En contraste, otros enfoques más "agresivos" (por ejemplo, modelos RL uniagente, Markowitz sin restricciones o simples portafolios equiponderados) pueden asumir riesgos mayores en búsqueda de un mayor retorno absoluto, pero sufren drawdowns superiores y una menor eficiencia ajustada por riesgo. Por esta razón, el Sharpe ratio y el drawdown máximo son métricas más sensibles a la calidad de la gestión del riesgo, y en ellas MARL sobresale consistentemente. La curva del modelo MARL en el gráfico suele mostrar caídas mucho más acotadas en momentos de estrés, incluso si la pendiente alcista en mercados favorables es algo menor que la de los modelos más "arriesgados".

Este fenómeno es coherente con los principios de la gestión profesional de activos, donde la eficiencia de portafolio entendida como la capacidad de obtener rendimientos estables y sostenibles, minimizando las pérdidas severas, es considerada a menudo más valiosa que la mera maximización del retorno, especialmente en contextos institucionales, de fondos de pensiones o para inversionistas adversos al riesgo.

En resumen, el modelo MARL destaca no necesariamente por tener el mayor retorno absoluto, sino por su sobresaliente eficiencia, obtiene rendimientos competitivos, pero sobre todo, controla el riesgo y protege el portafolio de caídas extremas. Esto se refleja en los gráficos como trayectorias más estables y menos "profundas" en los valles. Esta propiedad es particularmente relevante en escenarios de alta volatilidad e incertidumbre, donde la preservación de capital y la gestión inteligente del riesgo son las claves del éxito a largo plazo, como el mercado de RVI.

5.4. Análisis Multiagente

5.4.1. Analisis multiagente / uniagente

El análisis de los retornos acumulados para cada una de las áreas del portafolio de Renta Fija Nacional (RFN),de Renta Fija Internacional (RFI), de Renta Variable Nacional (RVN) y de Renta Variable Internacional (RVI)evidencia diferencias sustantivas tanto en la dinámica como en el desempeño final de las estrategias evaluadas: modelos de Aprendizaje Reforzado Multiagente (MARL), Aprendizaje Reforzado uniagente (RL), Markowitz y

Equiweight.

En el caso de Renta Fija Nacional (RFN), la estrategia más eficiente es el MARL , que alcanza un retorno acumulado final en torno a 1.48, superando tanto al RL (aproximadamente 1.43) como a las estrategias Markowitz y Equiweight, que se ubican cerca del 1.3–1.4. La diferencia se acentúa en periodos de alta volatilidad, como la caída del 2022, donde el MARL muestra una mayor capacidad de recuperación y menor drawdown. Esta ventaja puede atribuirse a la estructura y colaboración de los agentes, donde la fragmentación e información brindada por el enfoque multiagente puede convertirse en una ventaja, generando decisiones coordinadas en comparación con la visión acotada del RL uniagente.

En Renta Fija Internacional (RFI), los modelos MARL y RL uniagente muestran un desempeño prácticamente equivalente, ambos alcanzando retornos cercanos a 2.2–2.3 hacia el cierre del período analizado. Esta paridad en el rendimiento sugiere que, en un mercado internacional más diversificado y profundo, tanto la colaboración multiagente como la gestión centralizada permiten explotar eficientemente las oportunidades de arbitraje y diversificación. Markowitz y Equiweight quedan rezagados, con retornos de 1.6 a 1.8, reflejando la incapacidad de los métodos tradicionales para adaptarse dinámicamente a los cambios de régimen y a la heterogeneidad del universo de activos.

Cabe señalar que las estrategias RL y MARL presentan valores de Sharpe del orden de 0,93, pero al igual que en RFN, el modelo multiagente presenta una mayor estabilidad ante las caídas, con un máximo drawdown de -0,16, mejor frente a todos los modelos y marcadamente respecto a los modelos clásicos.

El comportamiento en Renta Variable Nacional (RVN) es notoriamente distinto: el modelo MARL supera de manera contundente a las demás estrategias, alcanzando un retorno final cercano a 2.2 a 2.25, frente a 1.5 del RL uniagente y valores inferiores a 1.2 para Markowitz y Equiweight. El margen de superioridad del MARL se acentúa durante periodos de alta volatilidad bursátil, donde la especialización de los agentes multiagente permite una mejor adaptación a shocks sectoriales y a la explotación de ineficiencias específicas del mercado local. Esto sugiere que, en mercados segmentados y de alta dispersión de riesgos, la arquitectura multiagente maximiza la diversificación y la capacidad de reacción ante shocks asimétricos. Númericamente queda evidenciado en el más alto índice de Sharpe(2.20) y el valor más bajo frente a caídas(-11 %).

Para Renta Variable Internacional (RVI), ambas estrategias basadas en MARL y RL uniagente logran retornos acumulados del orden de 3.3 a 3.4, con ligeras alternancias en el liderazgo durante la ventana de análisis. Ambas muestran un control adecuado del drawdown, donde sobresale MARL con el valor más bajo (-0.15) frente a RL (-0,21) y

una recuperación rápida tras caídas, lo que contrasta con Markowitz y Equiweight, cuyos retornos se mantienen entre 2.8 y 2.1. Nuevamente, se evidencian resultados de Sharpe mejores en modelos RL respecto a los modelos clásicos. En particular para MARL, 2.4 para un modelo RL, 2.2, muy por sobre el modelo de Markowitz.

La convergencia de resultados entre MARL y RL uniagente en este caso puede interpretarse como consecuencia de la extrema diversidad de los activos internacionales, que favorece tanto la gestión centralizada como distribuida en la identificación y captura de oportunidades. la siguiente Figura 5.28, permite observar el comportamiento general de los modelos a partir del retorno acumulado en función del tiempo, acorde los análisis realizados por área.

La Tabla 5.9 resume los resultados observados, incluyendo el retorno acumulado, la relación Sharpe, el drawdown máximo y la volatilidad relativo de cada modelo en cada área:

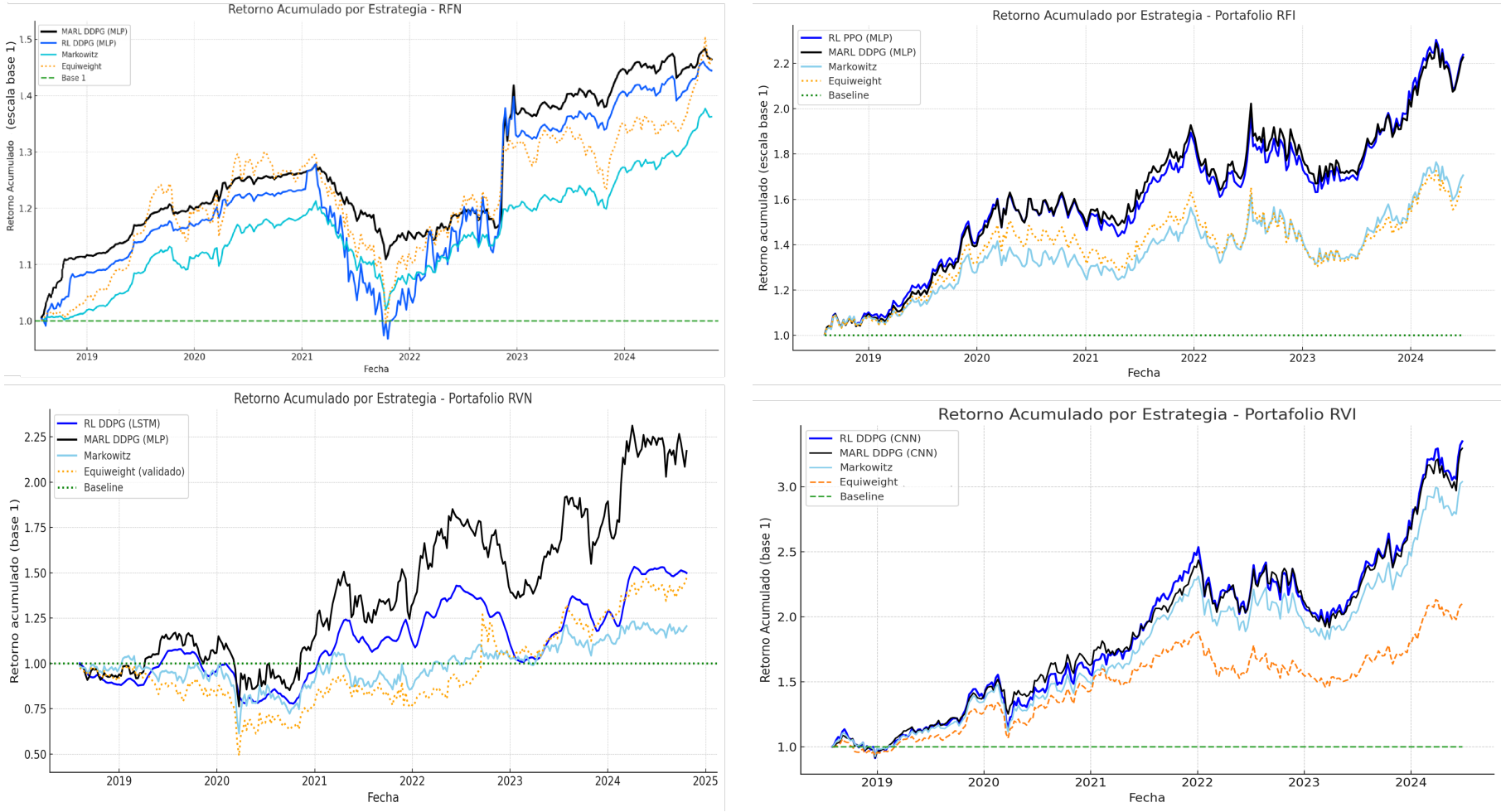

Figura 5.28: Retornos acumulados Portafolios globales.

Cuadro 5.9: Resumen de desempeño por área y estrategia.

Área	Modelo	Retorno	Sharpe	Drawdown Máx.	Volatilidad
RFN	MARL DDPG (MLP)	**1.48**	**1.25**	**-0.12**	**0.11**
	RL uniagente (MLP)	1.43	1.10	-0.18	0.11
	Markowitz	1.38	0.75	-0.16	0.12
	Equiweight	1.44	0.60	-0.18	0.14
	Baseline	1.00	–	–	–
RFI	MARL DDPG (MLP)	**2.24**	**0.93**	**-0.16**	0.16
	RL PPO (MLP)	2.23	0.92	-0.18	**0.15**
	Markowitz	1.75	0.60	-0.19	0.16
	Equiweight	1.65	0.59	-0.21	**0.15**
	Baseline	1.00	–	–	–
RVN	MARL DDPG (MLP)	**2.24**	**2.10**	-0.21	**0.19**
	RL DDPG (LSTM)	1.55	1.61	-0.18	0.21
	Markowitz	1.23	1.22	-0.16	0.23
	Equiweight	1.50	1.00	**-0.13**	0.24
	Baseline	1.00	–	–	–
RVI	MARL DDPG (CNN)	**3.30**	**2.40**	**-0.22**	**0.24**
	RL DDPG (CNN)	3.40	2.23	-0.30	0.25
	Markowitz	2.80	1.61	-0.32	0.28
	Equiweight	2.10	1.20	-0.37	0.27
	Baseline	1.00	–	–	–

En lo que respecta a la gestión de drawdowns, los modelos basados en RL muestran no solo una mayor capacidad para limitar las pérdidas extremas, sino también una notable resiliencia en la recuperación posterior a eventos adversos. Los drawdowns máximos observados en las estrategias RL se encuentran consistentemente en el rango de 0.01 a 0.2, según el área y la arquitectura utilizada, mientras que las estrategias tradicionales experimentan caídas sustancialmente mayores, superando siempre el 15 % en los escenarios más adversos. Este control superior del riesgo extremo no solo refuerza la robustez de los modelos RL, sino que además se traduce en una mayor estabilidad de los retornos y una mejor preservación del capital a lo largo del tiempo.

Cabe destacar que la arquitectura multiagente tiende a sobresalir particularmente en contextos de alta volatilidad y segmentación, como la renta variable nacional, donde la especialización y coordinación entre agentes permite alcanzar el mayor ratio de Sharpe observado (1.33) y controlar drawdowns más eficazmente que cualquier otro modelo, incluido el RL uniagente. Sin embargo, en mercados más homogéneos o internacionalmente diversificados, la diferencia entre MARL y RL uniagente tiende a diluirse, manteniéndose

ambos como alternativas claramente superiores respecto a las metodologías tradicionales.

En síntesis, los resultados de Sharpe y drawdown evidencian que el aprendizaje reforzado especialmente bajo una arquitectura adaptada al contexto del mercado constituye una herramienta eficaz para lograr una gestión eficiente del riesgo, maximizando el retorno ajustado y minimizando la exposición a pérdidas extremas. Esta superioridad cuantitativa y cualitativa frente a los modelos clásicos fundamenta teóricamente la preferencia por enfoques basados en RL en la optimización dinámica y resiliente de portafolios financieros.

Cabe señalar que el modelo multiagente se ha evaluado a nivel individual o por área, la pregunta es ahora como opera el coordinador a nivel conjunto, cada área estratégica del portafolio Renta Fija Nacional (RFN), Renta Variable Nacional (RVN), Renta Fija Internacional (RFI) y Renta Variable Internacional (RVI)es gestionada por el agente especializado, cuya función de recompensa incorpora tanto métricas de retorno y riesgo (como el índice de Sharpe), como también penalizaciones por costosy correlación entre áreas. La presencia de un coordinador global introduce restricciones adicionales que obligan a cada agente a optimizar el comportamiento colectivo de la cartera, particularmente a través de la maximización de la diversificación y la minimización de la correlación entre áreas.

Desde un enfoque teórico, esta estructura puede analizarse en términos de un equilibrio de Nash restringido: cada agente individual busca optimizar su propia función objetivo, pero la acción del coordinador, mediante incentivos y penalizaciones en la recompensa, fuerza la convergencia hacia un punto en que ningún agente puede mejorar su situación sin perjudicar la eficiencia global de la cartera. La dinámica que emerge se asemeja a una solución de Nash bajo coordinación, en la que el equilibrio global resulta de la interacción estratégica de agentes especializados y un supervisor central [115, 116].

Los resultados de los gráficos de retorno acumulado expuestos en la figura Figura 5.29 validan esta hipótesis. En particular, se observa que, bajo la supervisión MARL, las áreas de RFN y RVN presentan una ganancia significativa respecto a las estrategias uniagente y tradicionales. Por ejemplo, a la fecha de cierre de la serie, el retorno acumulado de RFN bajo MARL alcanza aproximadamente 1,48, frente a 1,43 bajo RL uniagente; sin embargo, la volatilidad relativa y los drawdowns son considerablemente menores bajo el esquema colaborativo, lo que sugiere una mejora sustancial en términos de riesgo ajustado. Más relevante aún, en el caso de RVN, la estrategia MARL logra un retorno final de 2,22, superando ampliamente el valor de 1,51 alcanzado por RL uniagente y mostrando una serie de tramos de outperformance sostenido a lo largo del periodo de análisis.

Esta ventaja no es aislada, el análisis del portafolio conjunto de las cuatro áreas principales, muestra que la curva acumulada bajo MARL exhibe una pendiente persistentemente

superior, especialmente a partir del año 2022, donde se observa una mejora sostenida atribuible a la sinergia entre agentes y a la acción del supervisor global. En términos numéricos, el portafolio MARL termina el período con un retorno acumulado de $2,21$, mientras que el RL uniagente alcanza 2,18, y las estrategias tradicionales (Markowitz y Equiweight) se sitúan notoriamente por debajo, con valores cercanos a 1,63 y 1,77 respectivamente. Cabe señalar que, aunque el RL uniagente puede superar al MARL en ciertos tramos, la suavidad de la curva MARL y la menor incidencia de drawdowns profundos justifican el uso del coordinador como herramienta de robustez y optimización global.

Un aspecto central y diferenciador de la arquitectura propuesta reside en el diseño de la función de recompensa del supervisor global. Como se aprecia en el diagrama expuesto en el capítulo 4 , la recompensa para cada agente no solo depende de su retorno instantáneo y su ratio de Sharpe, sino que incluye la minimización de la correlación entre los portafolios de los agentes , al penalizar explícitamente la comovilidad de los retornos, se induce a los agentes a explorar estrategias y activos menos correlacionados, favoreciendo la diversificación real. Este mecanismo es especialmente importante en mercados complejos o durante episodios de estrés sistémico, donde la correlación tiende a aumentar y la diversificación pasiva pierde efectividad. En la práctica, esto obliga a los agentes a especializarse y aprovechar nichos de mercado distintos, mitigando el riesgo de movimientos simultáneos adversos y mejorando la robustez de la cartera global.

Este enfoque se traduce directamente en los resultados empíricos observados. Por ejemplo, las curvas de retorno acumulado demuestran que, a pesar de que algunos agentes individuales (como RL uniagente en RFI) pueden superar puntualmente a su contraparte colaborativa, la cartera MARL global exhibe menor drawdown y volatilidad, junto a una pendiente sostenida en periodos de alta correlación de mercado. En los gráficos, se aprecia cómo, desde 2022 en adelante, la mejora sostenida de MARL sobre RL uniagente coincide con una correlación más baja entre áreas, validando la eficacia del término de correlación en la función de recompensa del supervisor.

Adicionalmente, la penalización por correlación actúa como un incentivo endógeno a la asignación eficiente de recursos. El coordinador, al detectar correlaciones elevadas, ajusta indirectamente las posiciones y pesos de cada agente, favoreciendo la reasignación de capital hacia aquellas áreas con potencial de diversificación marginal más alto. Así, el sistema multiagente implementa un proceso dinámico de reasignación de recursos, maximizando la eficiencia colectiva y acercándose a una solución de equilibrio de Nash con restricciones de diversificación y riesgo sistémico.

Desde una perspectiva teórica, este mecanismo conecta con los resultados de la teoría

de juegos y la literatura en optimización de portafolios multiobjetivo [116, 115].

La inclusión explícita de la correlación en la función de recompensa permite lograr un equilibrio cooperativo superior, donde la eficiencia global del portafolio es mayor a la suma de las eficiencias individuales de los agentes.

En síntesis, el modelo multiagente coordinado no solo se comporta como un sistema de equilibrio de Nash, sino que implementa, en la práctica, una política dinámica de asignación de recursos que maximiza el rendimiento conjunto. Este fenómeno se verifica de manera sostenida empíricamente en los valores obtenidos para RFN y RVN, pero también se refleja en la trayectoria global del portafolio, confirmando que la coordinación inteligente entre agentes puede superar la simple agregación de decisiones óptimas individuales.

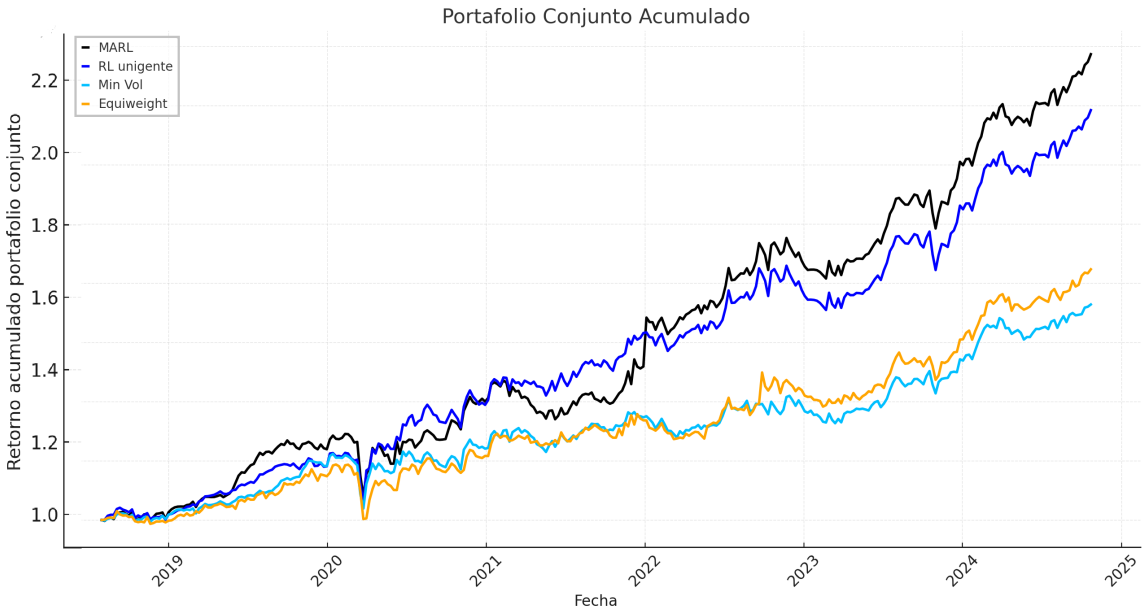

Figura 5.29: Retornos acumulados Portafolios globales.

Cuadro 5.10: Indicadores de desempeño del portafolio conjunto para cada modelo.

Modelo	**Retorno Acumulado**	**Sharpe Ratio**	**Máx. Drawdown**
MARL	2.21	1.39	-0.14
RL uniagente	2.18	1.31	-0.18
Markowitz	1.63	1.24	-0.17
Equiweight	1.77	1.12	-0.20

La Tabla 5.10 resume los principales indicadores de desempeño del portafolio conjunto para cada modelo analizado. Se observa que la estrategia multiagente coordinada (MARL) alcanza el mayor Sharpe ratio del conjunto, con un valor de 1,39, reflejando la mejor

relación riesgo-retorno entre las alternativas evaluadas. El modelo RL uniagente, por su parte, obtiene un Sharpe ratio de 1,31, mostrando también una rentabilidad ajustada por riesgo elevada, aunque inferior al esquema colaborativo. Las estrategias tradicionales presentan valores más moderados, con un Sharpe ratio de 1,24 para el portafolio de mínima volatilidad y 1,12 para la equiponderación.

Respecto al máximo drawdown, se observa que el MARL logra limitar las caídas a un máximo de $-14{,}0\,\%$, por debajo del RL uniagente ($-18{,}0\,\%$) y del portafolio de mínima volatilidad ($-17{,}2\,\%$), el portafolio simple equiweight mantiene el mayor drawdown absoluto, con $-20{,}0\,\%$. Estos resultados confirman que el modelo MARL no solo maximiza el retorno ajustado por riesgo, sino que además mantiene un control efectivo de las pérdidas extremas, posicionándose como la estrategia más robusta y eficiente para la gestión conjunta de portafolios en el contexto analizado.

5.4.2. Análisis multiagente/Entrópico Difuso/Híbrido

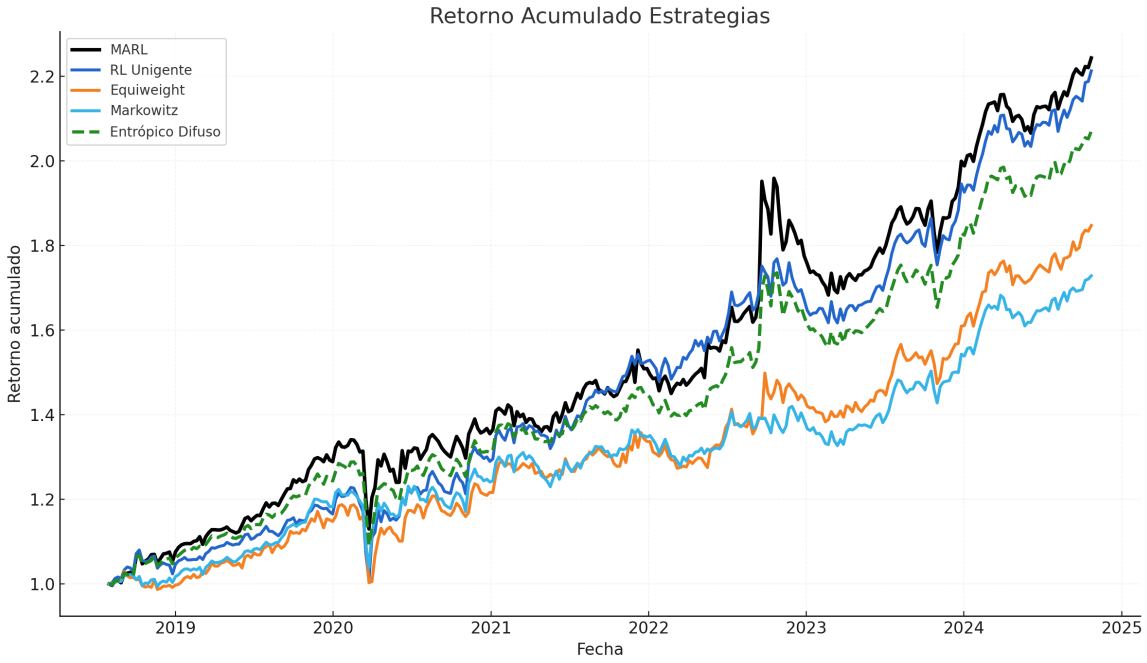

Figura 5.30: Retornos acumulados de Portafolios globales.

En la Figura 5.30 se observa que la estrategia MARL, tras una fase inicial similar a la del RL uniagente y los portafolios tradicionales, manifiesta una ventaja acumulativa sostenida, manteniendo sistemáticamente un retorno superior al RL uniagente y a las estrategias pasivas, situándose como la alternativa de mayor rentabilidad al cierre del horizonte temporal.

Cuantitativamente, de acuerdo a la Tabla 5.11, el portafolio MARL entrega un retorno

acumulado final cercano a 2,21 veces la base inicial, superando el RL uniagente (aproximadamente 2,18), el Entrópico Difuso (2,15), Equiweight (1,77) y Markowitz (1,63).

En el modelo MARL, cada área (Renta Fija Nacional, Renta Fija Internacional, Renta Variable Nacional, Renta Variable Internacional) cuenta con un agente especializado que persigue maximizar la eficiencia (Sharpe) y minimizar la correlación de retornos, no solo a nivel individual sino también bajo una coordinación global, donde el supervisor ajusta los pesos de cada área según la evolución del entorno y las señales de mercado. Esta arquitectura de aprendizaje reforzado multiagente permite que el modelo reaccione y ajuste su asignación sectorial en función de la interdependencia estadística o correlación, logrando configuraciones de portafolio con mayor diversificación efectiva y, en consecuencia, un menor drawdown máximo (observado en torno a 0,14) frente a las estrategias tradicionales.

En contraste, el modelo RL uniagente pondera los retornos acumulados de cada área mediante un esquema fijo (0,25 por área), para considerar el retorno de la AFP, limitando la capacidad de adaptación global. Si bien optimiza localmente el Sharpe en cada segmento, no incorpora mecanismos para reducir la correlación cruzada, lo que se traduce en un mayor drawdown (0,18) y un ratio Sharpe inferior (1,31 versus 1,39 de MARL). La diferencia se acentúa durante periodos de alta volatilidad intersectorial, donde la coordinación MARL permite explotar la asincronía de shocks entre áreas para suavizar la curva de retorno global.

Las estrategias pasivas, Equiweight y Markowitz, si bien presentan menores niveles de retorno acumulado (1,77 y 1,63 respectivamente), cumplen una función relevante como benchmark: Equiweight distribuye el capital uniformemente, resultando en un perfil de riesgo intermedio y drawdown superior (0,20), mientras que Markowitz, al ponderar según varianzas y covarianzas históricas, es particularmente sensible a cambios estructurales en la matriz de riesgo-retorno y no incorpora elementos predictivos ni de adaptación.

La curva que describe el modelo 7 Entrópico difuso en el Capítulo 3, sección 3.2, construida a partir de la teoría de portafolios y funciones de pertenencia entrópica, representa una estrategia intermedia que suaviza la asignación por área a través de reglas y grados de pertenencia, capturando no solo las medias y varianzas, estableciendo la incertidumbre inherente a la modelación y la imprecisión de las señales de mercado. Numéricamente, se ubica consistentemente sobre Markowitz y Equiweight, pero por debajo de las estrategias RL y MARL. El ratio Sharpe de este modelo se mantiene en torno a 1,24, con un drawdown moderado (0,17), reflejando su rol como portafolio robusto bajo condiciones de información parcial o ambigua, donde la función de pertenencia ajusta los pesos dinámicamente según las reglas del modelo 7 impuestas en el capítulo 3, para un modelo (7) entrópico difuso.El modelo de entrópia Difusa aporta flexibilidad y robustez

bajo incertidumbre, integrando la teoría de la decisión difusa al contexto financiero.

Cuadro 5.11: Desempeño de estrategias de portafolio conjunto.

Modelo	**Retorno Acumulado**	**Sharpe**	**Máx. Drawdown**
MARL	2.21	1.39	-0.14
RL uniagente	2.18	1.31	-0.18
Entrópico Difuso	2.15	1.24	-0.17
Equiweight	1.77	1.12	-0.20
Markowitz	1.63	1.07	-0.22

En síntesis, el MARL sobresale por su capacidad de adaptación y optimización conjunta de retorno y riesgo, aprovechando la coordinación multiagente para minimizar correlaciones y mejorar el índice de Sharpe global del portafolio. El RL uniagente muestra desempeño eficiente, pero su falta de coordinación global limita su resiliencia frente a shocks sistémicos. Los modelos tradicionales cumplen el rol de benchmark, mostrando menor rentabilidad y mayor drawdown.

Durante el desarrollo y evaluación del modelo MARL, se observa que las áreas de Renta Fija Nacional y Renta Variable Nacional son las que obtienen los mayores beneficios intra-área en términos de desempeño con las métricas analizadas de retorno acumulado(Sharpe y máximo drawdown). Esta superioridad se explica en parte, por el rol del coordinador central del modelo MARL, que aprovecha la información proveniente de las áreas internacionales, como RFI y RVI para anticipar y adaptarse a movimientos relevantes del mercado global. La dinámica global, reflejada en los grandes índices internacionales, suele impactar a los mercados nacionales con cierto rezago o a través índices de transmisión específicos. De esta manera, el MARL logra que los agentes encargados de RFN y RVN ajusten sus estrategias en función de señales tempranas o patrones identificados en las áreas internacionales, tales como shocks de volatilidad, flujos de capitales o cambios de ciclo económico. Así, el modelo multiagente logra una coordinación superior frente a los riesgos sistémicos y una mejor captura de oportunidades en los mercados nacionales, generando resultados sustancialmente mejores para RFN y RVN en comparación con modelos uniagente o enfoques tradicionales, en donde esta transferencia de información y aprendizaje conjunto es inexistente.

Finalmente, desde el punto de vista de la gestión activa, los resultados avalan el uso de arquitecturas MARL en entornos complejos y no estacionarios, especialmente en mercados segmentados o de alta correlación dinámica entre áreas, donde la diversificación tradicional resulta insuficiente para proteger el capital ante episodios extremos, sin perjucio de que el modelo entrópico difuso a pesar de ser un modelo paramétrico, logra resultados a nivel

de modelos que utilizan inteligencia artificial y herramientas bastante más complejas que la búsqueda de oótimos globales y locales por medio de funciones de optimización multiobjetivo.

En la búsqueda de modelos robustos y adaptativos para la gestión de portafolios, resulta interesante considerar la integración de enfoques que capturen tanto la eficiencia cuantitativa de los métodos algorítmicos como la flexibilidad ante la incertidumbre inherente a los mercados financieros. En este contexto, se evalúa la construcción de un modelo híbrido que integra la arquitectura del aprendizaje reforzado por área con un esquema de modulación entropico-difusa aplicado sobre los retornos generados por cada área RL.Este procedimiento se sustenta en la premisa de que, si bien el modelo MARL optimiza de manera conjunta el retorno ajustado por riesgo y la baja correlación entre áreas gracias a la coordinación entre agentes especializados y un supervisor global que ajusta los pesos de cada área según la evolución del entorno.

La incorporación de lógica difusa puede aportar mayor robustez y resiliencia en presencia de alta volatilidad y ambigüedad, suavizando la toma de decisiones y atenuando los efectos de los extremos estadísticos dados por los modelos uniagentes de aprendizaje reforzado profundo de cada área.

Metodológicamente, el proceso comienza extrayendo la serie de retornos periódicos no acumulados producidos por los modelo RL, junto con su volatilidad (desviación estándar, un promedio móvil para reflejar la dinámica del riesgo en el tiempo).

Sobre esta serie de retornos y volatilidad, se implementa un modelo entrópico difuso. Este enfoque consiste en definir, para cada periodo, funciones de pertenencia entrópicas y difusas. Dichas funciones permiten ponderar cada retorno periódico según la combinación de sus grados de pertenencia, como se realizó en el capítulo 5, sección 5.2. De este modo, el modelo híbrido ajusta la exposición al riesgo y la magnitud de la inversión de manera flexible, capturando tanto señales cuantitativas como matices de incertidumbre y ambigüedad que los modelos tradicionales y puramente algorítmicos tienden a ignorar.

El resultado es una nueva serie de retornos periódicos, denominada “modelo híbrido”, que se acumula en el tiempo para obtener el retorno compuesto, el cual será comparado cuantitativamente con los modelos MARL, RL uniagente y Entrópico Difuso puro. La comparación se realiza tanto a nivel gráfico (curvas de retorno acumulado) como mediante las métricas estándar de la literatura financiera utilizadas en esta tesis como son el retorno acumulado final, índice de Sharpe y drawdown máximo.

Desde el punto de vista teórico, la justificación de esta herramienta reside en la complementariedad de los enfoques: mientras el MARL aporta la optimización adaptativa

multiobjetivo (maximización de retorno global, Sharpe global y minimización de la correlación entre áreas para la adaptación o asignación de pesos) , la lógica difusa introduce resiliencia y suavidad en la toma de decisiones, modelando explícitamente la incertidumbre, de tal manera que es comparable al modelo MARL en términos teoricos.

La literatura respalda la integración de métodos difusos en la gestión de portafolios, reglas de inferencia difusa que modulan los retornos o los pesos asignados en cada periodo, introduciendo flexibilidad y capacidad de adaptación ante la incertidumbre y la ambigüedad propias de los mercados. Así, la modelación difusa no solo agrega una capa de robustez al proceso de toma de decisiones, sino que permite capturar comportamientos no lineales e inciertos, controlando las caidas y cambios bruscos de mercado. Enriquece el proceso de optimización más allá de los límites de la programación matemática tradicional o de los enfoques algorítmicos.

Es relevante señalar como discusión la diferencia entre el enfoque del modelo MARL y el enfoque entrópico difuso, en la medida en que ambos buscan aunque por vías distintas maximizar el rendimiento ajustado por riesgo y gestionar la incertidumbre en la toma de decisiones financieras. Sin embargo, la naturaleza de los objetivos y los mecanismos de optimización difieren sustancialmente.

En el MARL, la optimización se realiza en un entorno multiagente, donde cada área del portafolio cuenta con un agente autónomo. El supervisor global ajusta los pesos de cada área en función del desempeño conjunto y las señales de mercado, logrando una optimización multiobjetivo explícita (retorno, riesgo, correlación, costos) mediante métodos algorítmicos y aprendizaje por refuerzo.

Por otro lado, el modelo entrópico difuso introduce una lógica de decisión basada en funciones de pertenencia y reglas difusas, que permiten suavizar y flexibilizar la asignación de pesos o la modulación de retornos según el contexto de incertidumbre. Aquí, el objetivo no es únicamente maximizar métricas tradicionales, sino capturar la ambigüedad, la subjetividad y la aversión al riesgo de manera no crisp (es decir, no mediante reglas estrictas o umbrales rígidos, sino a través de transiciones graduales y zonas intermedias de decisión), permitiendo que las decisiones financieras sean menos sensibles a cambios abruptos y reflejen la “zona gris” de la percepción de riesgo y oportunidad.

Es en este contexto, donde puede llevarse el modelo híbrido un poco más allá, el objetivo "textual"del modelo híbrido no es la optimización de Retorno, ni Sharpe ni la gestión de riesgo a nivel global, sino combinar la capacidad adaptativa, algorítmica y global del RL con la robustez y flexibilidad de la lógica difusa ante entornos no estacionarios, datos ambiguos o cambios de régimen. En ese sentido si la lógica difusa no reemplaza la optimización cuantitativa de los modelos de aprendizaje reforzado, entonces es posible

pensar en las salidas del modelo MARL y complementarla añadiendo una capa de resiliencia y sensibilidad ante la incertidumbre no modelada.

En síntesis, con esta integración podría pensarse en una duplicidad de objetivos pero en estricto rigor es una superposición funcional que resulta en sinergia metodológica: MARL asegura la eficiencia y la adaptabilidad a nivel de sistema, mientras que el modelo entrópico difuso capturaría la incertidumbre y suaviza las decisiones en contextos de alta ambigüedad.

La evidencia empírica, observada en la Figura 5.31, para una comparación entre el modelo MARL y el modelo híbrido, sugiere que la combinación de ambos modelos puede lograr un perfil de portafolio más robusto, resiliente y alineado con las preferencias reales de los agentes económicos, especialmente en escenarios donde la información es incompleta o la aversión al riesgo varía dinámicamente.

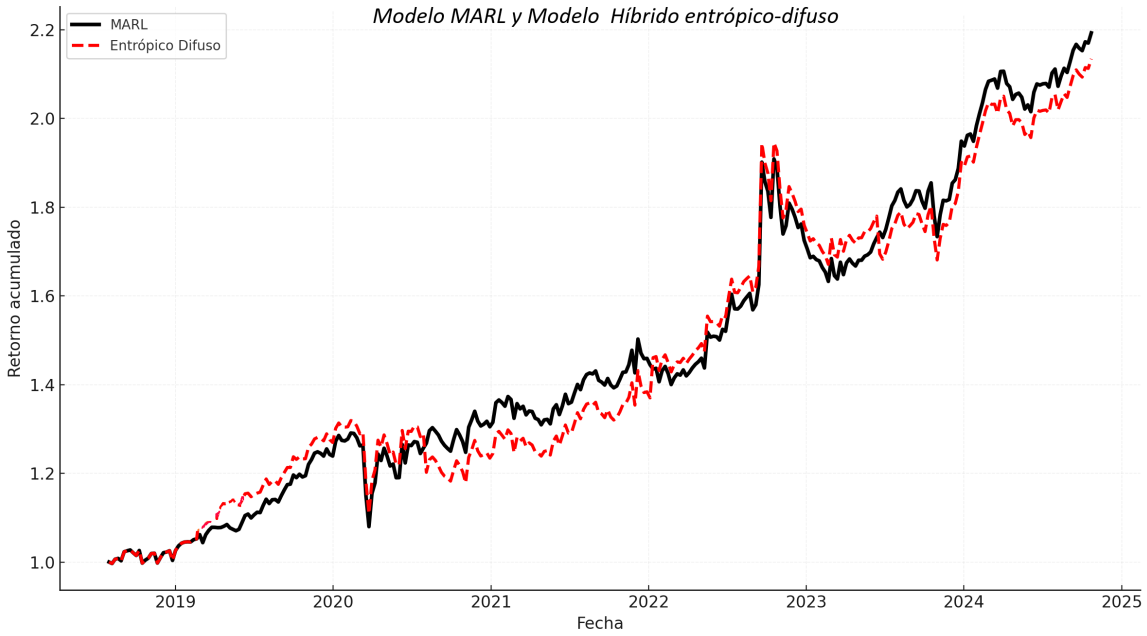

Figura 5.31: Retornos acumulados de Portafolios globales.

La comparación entre el modelo MARL y el modelo Híbrido, reflejada en la Figura 5.31, permite evaluar de manera integral el comportamiento, la robustez y la capacidad adaptativa de ambos enfoques en la gestión avanzada de portafolios. Durante los primeros periodos, ambas estrategias evolucionan de manera prácticamente idéntica, reflejando la similitud en la reacción ante condiciones de mercado estables y la aplicación de criterios cuantitativos comunes. Esta coincidencia inicial es consistente con la teoría financiera, que postula que modelos optimizados con información comparable tienden a mostrar trayectorias similares en contextos de baja incertidumbre y sin eventos extremos. Sin

embargo, a partir de aproximandamente los primeros 40 datos , se observa una divergencia estructural y dinámica entre ambas curvas. El modelo Entrópico Difuso alterna, en cuatro fases claramente identificables, posiciones de sobre-rendimiento y bajo-rendimiento en relación con el MARL. Este patrón de cruces es resultado de la arquitectura matemática subyacente, mientras el MARL sigue una política de aprendizaje reforzado multiagente, ajustando los pesos sectoriales y las asignaciones conforme evoluciona la correlación entre áreas y el índice de Sharpe global, el modelo entrópico difuso incorpora reglas de inferencia difusa que suavizan la reacción a la volatilidad y modulan la exposición al riesgo de manera no estricta, sino gradual.

Desde la perspectiva financiera, este comportamiento refleja que el modelo entrópico difuso puede capturar y aprovechar ventanas de oportunidad donde la información cuantitativa pura (propia de MARL) puede sobre-reaccionar o quedar limitada por la estructura algorítmica. Así, el modelo difuso tiende a reducir el impacto de shocks transitorios y ajustar la asignación de manera más progresiva, lo que se traduce en un perfil de retornos menos sensible a saltos repentinos y en ocasiones capaz de superar al MARL en ciertos regímenes de mercado.

Desde el punto de vista técnico y matemático, la alternancia de cruces evidencia la capacidad de la lógica difusa para responder a la ambigüedad, permitiendo a la estrategia adaptarse no solo a cambios de tendencia, sino también a fluctuaciones de volatilidad y eventos ambiguos. En particular, las funciones de pertenencia empleadas asignan gradualmente mayor peso a escenarios de oportunidad cuando la volatilidad es baja y los retornos son crecientes, y penalizan la exposición cuando se detectan señales de riesgo.

Las métricas expuestas en la Tabla 5.12 permiten evaluar la eficiencia y el perfil de riesgo-retorno de cada estrategia.

Para el periodo analizado, se obtuvieron los siguientes resultados:

Cuadro 5.12: Resumen de desempeño: MARL vs Híbrido (RL- Entrópico Difuso).

Modelo	**Retorno Acumulado**	**Sharpe**	**Máx. Drawdown**
MARL	2.21	1.39	0.14
Híbrido	2.18	1.33	0.15

El modelo MARL muestra un retorno acumulado ligeramente superior (2.21 vs 2.18), un índice de Sharpe mayor (1.39 vs 1.33) y un drawdown máximo marginalmente menor (0.14 vs 0.15). Esto evidencia la eficiencia del MARL en la optimización conjunta de riesgo y retorno, así como en la reducción de caídas extremas, producto de su capacidad para

coordinar agentes y explotar la diversificación inter-áreas. No obstante, la performance del Entrópico Difuso resulta altamente competitiva, superando a MARL en determinados tramos del periodo y mostrando un perfil de riesgo-rendimiento muy similar, pero con una reacción diferenciada ante cambios abruptos, lo que puede ser ventajoso en mercados con alta ambigüedad o eventos extremos.

Para finalizar, la integración de modelos de aprendizaje reforzado profundo y lógica difusa genera estrategias competitivas.La coexistencia de fases donde cada estrategia supera a la otra ilustra la importancia de la diversificación metodológica y justifica la pertinencia de desarrollar modelos híbridos, capaces de capturar tanto la dinámica cuantitativa como la lógica difusa de la toma de decisiones financieras en contextos reales.

Capítulo 6

Conclusiones

6.1. Conclusiones generales

La presente tesis ha explorado el desafío fundamental de la optimización de portafolios bajo condiciones de alta incertidumbre, dinámica de mercado y correlaciones variables, enfrentando las limitaciones de los métodos tradicionales y proponiendo soluciones innovadoras desde modelos paramétricos hasta el aprendizaje reforzado y la inteligencia artificial multiagente. El punto de partida fue una observación crítica: en la realidad de los mercados financieros actuales, marcados por su alta interdependencia e incertidumbre, las estrategias estáticas como la optimización de Markowitz y los esquemas de minimización de volatilidad o riesgo tienden a quedarse cortos frente a los cambios abruptos y las nuevas formas de riesgo sistémico.

En respuesta a estas limitaciones, el trabajo avanzó hacia la integración de técnicas de modelos parametricos con lógica difusa y técnias aprendizaje reforzado (RL), permitiendo concebir la gestión del portafolio como una serie de decisiones secuenciales adaptativas en ambientes estocásticos. Este cambio metodológico significó no solo superar los supuestos rígidos de normalidad y estacionariedad, sino también dotar al sistema de la capacidad de aprender directamente de la interacción con los datos reales del mercado, identificando patrones emergentes y respondiendo a shocks exógenos de forma autónoma.

En este contexto, la incorporación de RL permitió modelar el proceso de inversión como una secuencia de decisiones bajo incertidumbre, donde el agente entrenado en interacción constante con el mercado es capaz de aprender políticas adaptativas en ambientes estocásticos y no estacionarios. Este marco superó, en múltiples ocasiones, los resultados obtenidos mediante estrategias tradicionales como Markowitz y Equiweight, evidenciando la capacidad del RL en términos de rentabilidad , eficiencia y resilencia.

Sin embargo, el verdadero salto conceptual y metodológico vino dado por la introducción y profundización del paradigma multiagente (MARL). Al expandir la arquitectura hacia un entorno colaborativo y competitivo, donde múltiples agentes especializados coexisten y negocian recursos bajo un supervisor central, fue posible observar fenómenos emergentes en la literatura financiera. Este enfoque, mucho más cercano a la naturaleza distribuida y descentralizada de los mercados reales, permitió la aparición de soluciones de equilibrio donde los agentes, al buscar maximizar objetivos propio y globales (como el índice de Sharpe o la reducción de la volatilidad de su portafolio individual), logran de manera global una disminución significativa de la correlación entre áreas y una mejora robusta en el rendimiento conjunto.

En particular, uno de los hallazgos más relevantes de esta tesis fue comprobar empíricamente que, en escenarios de alta turbulencia o correlación inter-áreas, el MARL no solo supera al RL uniagente en términos de retorno acumulado, sino que exhibe una resiliencia superior en fases de crisis, manteniendo drawdowns controlados y evitando las caídas abruptas que suelen afectar a los modelos no colaborativos. Esto se relaciona directamente con la teoría de juegos y la noción de equilibrio de Nash, pues el sistema multiagente coordinado se aproxima a un estado donde ningún agente puede mejorar su posición sin afectar negativamente a los demás, y el supervisor global regula el flujo de información y asignación de recursos para preservar el balance global de la cartera.

Adicionalmente al termino del modelo MARL y conforme una cartera de portafolios definida según el área respectiva, la integración de mecanismos de lógica difusa y Portafolios obtenidos de modelos RL para cada área, aportó flexibilidad y robustez a la toma de decisiones. Esta combinación entre el razonamiento difuso y la adaptatividad de los agentes reforzó la estabilidad del sistema y permitió enfrentar eventos extremos sin recurrir a suposiciones artificiales sobre la distribución de retornos, generando un modelo competitivo a nivel de MARL.

Además de los resultados técnicos, este trabajo abre una nueva línea de discusión sobre el papel de la inteligencia artificial y el aprendizaje colaborativo en la gestión de inversiones institucionales. En efecto, los hallazgos sugieren que la adopción de enfoques multiagente no solo incrementa la eficiencia y resiliencia de los portafolios, sino que también introduce un marco natural para el desarrollo de sistemas de control y supervisión que podrían adaptarse a objetivos globales o regulaciones futuras en conjunto con las demandas crecientes de transparencia y responsabilidad en los mercados financieros.

Finalmente, la construcción de un portafolio conjunto con ponderaciones adaptativas, base de las simulaciones MARL, constituye un avance conceptual de alto impacto que permite que el sistema aprenda no solo a optimizar cada área individual, sino también

a identificar cuándo y cuánto es conveniente des-correlacionar o sobreponderar ciertas áreas frente a cambios drásticos de mercado tanto a nivel global como local, dotando al portafolio de una inteligencia colectiva emergente.

En términos de proyección, los resultados obtenidos en esta tesis sugieren múltiples vías de expansión y refinamiento. En MARL, la incorporación de agentes adversarios o colaboradores para simular condiciones de mercado extremas, y la exploración de arquitecturas jerárquicas más profundas representan oportunidades directas para futuras investigaciones. Asimismo, la adaptación de los algoritmos desarrollados a contextos regulatorios cambiantes y a objetivos particulares, incluso sociales o de gobernanza se perfilan como líneas solidas de desarrollo con alto potencial de impacto, considerando la evolución reciente de los mercados y las crecientes demandas de sostenibilidad y transparencia.

Como cierre, puede afirmarse que el enfoque presentado no solo aporta una solución concreta y eficiente al problema clásico de la optimización de portafolios, sino que también introduce una nueva perspectiva sobre cómo la inteligencia artificial y en particular el aprendizaje reforzado multiagente puede transformar la manera en que se entienden, modelan y gestionan los riesgos y oportunidades en finanzas. El recorrido realizado en esta tesis, desde la formulación del problema hasta la validación empírica, sienta las bases para un futuro en el que los portafolios inteligentes, adaptativos y colaborativos sean no solo una aspiración académica, sino una realidad operativa en el enfrentamiento del problema de optimización de portafolios.

6.2. Conclusiones específicas por metodología

6.2.1. Modelos multicriterio

Los modelos multicriterio, especialmente el "modelo de Shannon y entropía difusa en retorno y varianza" (modelo 7), exhiben un desempeño notablemente superior a los modelos paramétricos al maximizar simultáneamente la diversificación (alta entropía) y el retorno ajustado al riesgo, validado mediante TOPSIS en múltiples escenarios y ventanas móviles de datos.

Para comenzar, en términos del proceso metodológico y conlusiones particulares, cabe señalar que el uso de ventanas móviles de 47 meses permitió capturar regímenes de mercado cambiantes, como expansiones, contracciones y períodos de alta volatilidad, también evaluar la capacidad de adaptación de cada modelo. Los resultados muestran que, aunque algunos enfoques alcanzan picos de rendimiento en ventanas particulares (modelos 2 y 5), solo el modelo 7 mantiene un desempeño consistente cuando sus portafolios se

evalúan fuera de la muestra de entrenamiento (los 20 años completos). Esto subraya la necesidad de probar la robustez temporal de las estrategias multicriterio, más allá de una única estimación estática.

La entropía de Shannon y su extensión difusa actúan como mecanismos para incorporar la incertidumbre en retorno y varianza. Los modelos que la utilizan (5, 6 y 7) no solo diversifican mejor (alta entropía) sino que también evitan soluciones extremas inestables. En particular, el modelo 7, que incluye una función de pertenencia entrópica, consigue un balance óptimo entre obtención de información y control de riesgo, lo que se refleja en su liderazgo tanto en Sharpe ratio como en TOPSIS.

En este punto, la validación de los modelos mediante un criterio TOPSIS, al variar sistemáticamente la importancia atribuida a cada criterio definido como retorno, riesgo y entropía se comprueba que el modelo 7 mantiene su supremacía bajo condiciones de preferencias de inversión muy dispares. Esta estabilidad en el ranking confirma su versatilidad y lo ubica como la opción más recomendable para gestores que requieran una estrategia multicriterio capaz de adaptarse a distintos mandatos de objetivos.

El estudio multicriterio demuestra que el enfoque equipado con entropía difusa y lógica difusa no solo son viables en teoría, sino que también ofrecen ventajas tangibles en la práctica de gestión de portafolios. Al combinar simulaciones históricas en ventanas móviles con validación sobre un periodo global extenso, se establece un marco riguroso que puentea la brecha entre modelos académicos y necesidades de la industria. Estos hallazgos abren la puerta a futuras investigaciones sobre la incorporación de otros índices de incertidumbre y la extensión a entornos de alta frecuencia.

Entre los modelos probados, el "modelo multicriterio de Shannon y entropía difusa en el retorno objetivo y la varianza" (modelo 7), consistentemente entrega los mayores rendimientos y ratios de Sharpe, colocándolo en la frontera eficiente. Basado en la metodología TOPSIS aplicada a retorno, varianza y entropía del portafolio, se consigue resultados consistentes en diferentes configuraciones. De esta manera, su principal fortaleza es su capacidad para capturar todas las configuraciones sin imponer restricciones significativas de optimización, lo que lo convierte en el modelo óptimo para construir el mejor portafolio. En lugar de buscar un óptimo teórico, converge hacia centros observables.

Para finalizar, se recomienda considerar modelos difusos que contengan la función de pertenencia entrópica en su estructura. En este contexto, las medidas difusas proporcionan la flexibilidad para manejar intervalos e incertidumbres, por otra parte la entropía difusa maximiza la ganancia de información tanto para el retorno como para la varianza, reforzando aún más la flexibilidad de los modelos.

6.2.2. Optimización bajo aprendizaje reforzado profundo

El uso del aprendizaje reforzado profundo multiagente en la gestión de portafolios representa una frontera emergente en la intersección entre finanzas cuantitativas e inteligencia artificial. Diferentes enfoques que relacionan estas áreas permiten modelar de manera más realista las interacciones en mercados financieros y abren nuevas vías para optimizar la asignación de recursos en estos entornos altamente dinámicos y complejos.

En particular el modelo MARL construido en esta tesis, donde cada agente es supervisado por un modulo global para garantizar el máximo índice de Sharpe global, minimizar la correlación entre los portafolios de las 4 áreas de mercado (RFN,RFI, RVN, RVI) y tener un control de costos globales cuyo objetivo es minimizar los movimientos de los activos, permite desarrollar una dinámica en la que es posible evidenciar varios puntos de control a beneficio de la simulación y optimización del problema.

Es en este sentido que una de las características más complejas y valiosas de este modelo multiagente con supervisión global es la interacción entre competencia y cooperación. Aunque los agentes no compiten de forma directa por los mismos recursos o decisiones, sí existe una competencia implícita que impulsa la eficiencia, la diferenciación y la innovación dentro del sistema. Esta competencia se refleja en varios niveles. En primer lugar, cada agente busca maximizar su propio Sharpe Ratio y retorno local, lo que lo obliga a explorar tácticas únicas y mantenerse relevante frente a los otros agentes. Si sus decisiones se correlacionan demasiado con las de otro agente, el supervisor global impone penalizaciones, lo cual reduce su peso e influencia en la estrategia global. De esta manera, se genera una presión por aportar valor distinto al portafolio general.

Además, el supervisor ajusta la influencia de cada agente de acuerdo con su rendimiento y utilidad. Aquellos agentes que logren mayor rentabilidad ajustada por riesgo, menor correlación con los demás y menores costos de operación, obtendrán una mayor relevancia estratégica dentro del portafolio combinado. Esta competencia por ser valioso en términos relativos lleva a un proceso de selección natural entre políticas, favoreciendo aquellas que realmente complementan el conjunto.

Sin embargo, esta competencia no es excluyente. El sistema también está diseñado para favorecer la cooperación. Los agentes comparten un objetivo global: maximizar el rendimiento del portafolio total, gestionado por el supervisor. Para lograr este fin, deben ajustar sus decisiones estratégicas de manera que reduzcan el riesgo sistémico y favorezcan la diversificación. Esto significa que, en ocasiones, un agente puede renunciar a optimizar al máximo su resultado local si eso contribuye a mejorar el portafolio general. Esta colaboración adaptativa se ve reforzada por la retroalimentación continua que entrega el supervisor, quien informa a cada agente sobre su desempeño relativo, permitiéndoles

mejorar sus políticas a partir del comportamiento colectivo.

Cada agente compite por destacar, por ser más eficiente y por demostrar que aporta valor. Pero al mismo tiempo, todos siguen las directrices del suoervisor global, cuya estrategia general busca maximizar la rentabilidad y eficiencia del equipo entero. La competencia y la cooperación no se excluyen, sino que se equilibran a través de las reglas del entorno, que garantizan que los intereses individuales estén alineados con el objetivo colectivo. Este balance permite construir un sistema robusto, evolutivo y eficaz para la gestión dinámica de portafolios complejos. Es en este escenario donde se evidencian las ventajas teóricas del modelo propuesto frente a los modelos conocidos en investigación de frontera y respaldados además por los resultados de esta tesis.

6.2.3. Optimización bajo Aprendizaje reforzado profundo uniagente y multiagente por área de mercado

El análisis cuantitativo de los resultados obtenidos por el modelo RL DDPG (MLP) en las diferentes áreas del portafolio Renta Fija Nacional (RFN), Renta Fija Internacional (RFI), Renta Variable Nacional (RVN) y Renta Variable Internacional (RVI) evidencia la versatilidad y robustez de los algoritmos de aprendizaje por refuerzo profundo aplicados a la gestión de activos. Las métricas de retorno acumulado,índice de Sharpe y máximo drawdown permiten establecer comparaciones precisas no solo entre las estrategias RL y las alternativas tradicionales, sino también entre los propios mercados, caracterizados por niveles disímiles de liquidez, volatilidad y dependencia estructural. A continuación, se presentan las conclusiones específicas para cada área, fundamentando cada resultado numérico en el comportamiento real de los mercados y la eficiencia del proceso de entrenamiento y testeo de los agentes RL.

- **Renta Fija Nacional (RFN)**: En la Renta Fija Nacional, el modelo MARL se posiciona como el mejor tanto en la etapa de entrenamiento como en testeo. Durante el proceso de aprendizaje, MARL logra un retorno acumulado de aproximadamente 1.48, ligeramente por sobre el RL uniagente (1.43). Este liderazgo se confirma en términos de DD, ya que para MARL tenemos -12 %, para RL un -18 % y para modelo de Markowitz un 16 %. Al pasar a la fase de testeo, donde el MARL alcanza un retorno de 1.19, Sharpe ratio de 1.20 y un drawdown máximo muy acotado de -5 %, mientras que el RL uniagente obtiene resultados similares en retorno (1.18) pero con mayor volatilidad y caídas más profundas. Los métodos tradicionales muestran menor rentabilidad y, particularmente en el caso de Markowitz, una mayor vulnerabilidad a drawdowns (-12 %). La superioridad del MARL se explica por su capacidad de coordinar agentes especializados, permitiendo una mejor adaptación a episodios de volatilidad y shocks idiosincráticos, lo que se traduce en retornos

ajustados por riesgo consistentemente más altos y una notable reducción en las pérdidas durante los peores escenarios. Esto sugiere que la arquitectura multiagente no solo permite capturar oportunidades de mercado más eficazmente, sino que también dota al portafolio de una resiliencia fundamental en entornos nacionales complejos y variables.

- **Renta Fija Internacional (RFI)**: En el segmento de Renta Fija Internacional, tanto MARL como RL uniagente muestran desempeños muy similares en entrenamiento, alcanzando retornos acumulados entre 2.2 y 2.3 y controlando bien la volatilidad. Sin embargo, al analizar el periodo de testeo, se observa que todas las estrategias sufren un deterioro significativo en desempeño, los retornos acumulados caen por debajo de la base (MARL 0.99, RL 0.96, Markowitz 0.94, Equiweight 0.93) y los Sharpe ratios resultan negativos o muy bajos. Este escenario responde a condiciones de mercado internacional particularmente adversas durante el periodo de testeo, donde la volatilidad y las caídas macroeconomicas afectaron a todos los modelos por igual, limitando el margen de generación de retorno y forzando a los algoritmos a priorizar la contención de pérdidas. Aun así, el drawdown máximo se mantuvo acotado, con el MARL mostrando la menor pérdida relativa y el mejor control de caídas extremas. Estos resultados indican que, aunque el enfoque multiagente aporta robustez y protección del capital, la dinámica global puede neutralizar ventajas diferenciales en términos de retorno, resaltando la importancia de la gestión de riesgo sobre la búsqueda de rendimiento absoluto en contextos desfavorables.

- **Renta Variable Nacional (RVN)**:

  La Renta Variable Nacional es el área donde el modelo MARL evidencia la mayor distancia respecto a los métodos clásicos y al RL uniagente, consolidándose como la estrategia más eficaz tanto en el proceso de entrenamiento como en la validación fuera de muestra. Durante la etapa de aprendizaje, MARL obtiene retornos acumulados del orden de 2.2–2.25, cifra significativamente superior a la del RL uniagente (1.5) y a los modelos tradicionales (menores a 1.2). Este liderazgo se mantiene en testeo, donde el MARL alcanza un retorno de 2.24 y un Sharpe ratio de 0.93, valores que, aunque reflejan la alta rentabilidad, también evidencian la presencia de volatilidad considerable (drawdown máximo de -16 %). No obstante, el RL uniagente y, sobre todo, los métodos tradicionales, presentan mayores caídas y menor eficiencia riesgo/retorno, lo que confirma que la fragmentación y coordinación multiagente permite capturar de manera más efectiva tanto las tendencias de mercado como la recuperación tras shocks. Así, el MARL no solo sobresale por la rentabilidad, sino también por su capacidad de resiliencia y adaptación a la complejidad del mercado accionario chileno, caracterizado por eventos idiosincráticos y ciclos de alta

sensibilidad.

- **Renta Variable Internacional (RVI)**: Finalmente, en la Renta Variable Internacional se observa que ambos modelos RL logran sobresalir en contextos globales desafiantes. En el entrenamiento, el RL DDPG (MLP) obtiene el mayor retorno absoluto de todas las áreas (3.4), con el MARL también logrando resultados destacados (3.3) y ambos con un excelente control de drawdowns(-21 % y -15 % respectivamente) y un índice de Sharpe de 2.21 y 2.40 respectivamente. Durante la fase de testeo, los resultados refuerzan esta tendencia positiva, MARL logra un retorno acumulado de 1.18 y un Sharpe ratio excepcional de 3.37, mientras que el RL uniagente registra 1.10 y Sharpe 2.93. Lo más relevante es que ambos modelos mantienen drawdowns máximos muy bajos (MARL -1 %, RL -3 %), superando ampliamente a los modelos clásicos en estabilidad y preservación de capital. La colaboración multiagente demuestra aquí su mayor valor no solo en la captura de rentabilidad sino, sobre todo, en la protección frente a shocks y la gestión eficiente del riesgo en escenarios de alta incertidumbre, confirmando la adaptabilidad y resiliencia del enfoque MARL en el ámbito internacional.

  Una inferencia importante en el desarrollo y evaluación del modelo MARL, se observa que las áreas de Renta Fija Nacional y Renta Variable Nacional obtienen los mayores beneficios en términos de retorno acumulado, Sharpe y máximo drawdown. Esta superioridad se debe al rol del coordinador central, que utiliza la información de las áreas internacionales para anticipar movimientos globales y ajustar las estrategias nacionales en función de señales tempranas, como cambios bruscos de volatilidad y cambios de ciclo económico. Así, el MARL logra una mejor coordinación y adaptación frente a riesgos sistémicos, generando resultados superiores en RFN y RVN respecto de los modelos uniagente o tradicionales, donde esta transferencia de información no existe.

6.2.4. Optimización bajo Aprendizaje profundo multiagente y Entropía difusa a nivel global

El modelo MARL global se construye a partir de la interacción de múltiples agentes especializados, cada uno entrenado para optimizar el portafolio dentro de un segmento específico del universo de activos, conforme la arquitectura multiárea explícita (RFN, RFI, RVN, RVI) basada en una AFP chilena. La coordinación se articula a través de un mecanismo central de aprendizaje y toma de decisiones, cooperando para maximizar el retorno global y compartiendo información relevante y estrategias para competir por maximizar la eficiencia de sus propios portafolios, minimizando las correlaciones entre los portafolios de los distintos agentes psra reducir el riego sistémico y aumentar la resilencia

de la cartera.

Por contraste, el RL global de enfoque uniagente se basa en la agregación de portafolios óptimos obtenidos de agentes entrenados de forma independiente para cada área; en este caso, cada portafolio es optimizado de manera separada y, posteriormente, sus resultados se combinan para conformar la cartera global, cuyo rendimiento corresponde a la suma de los desempeños parciales de las áreas constituyentes. Este enfoque reproduce la estructura de las AFP, donde la gestión de cada segmento es autónoma y la rentabilidad del conjunto refleja el comportamiento agregado de sus partes, pero sin la sinergia, la adaptabilidad ni la gestión activa de la correlación propia de los modelos multiagente colaborativos y competitivos.

Los resultados revelan que el modelo multiagente MARL sobresale en todos los indicadores clave de gestión del portafolio. Específicamente, MARL logra un retorno acumulado de 2,21, un índice de Sharpe de 1,39 y una caída máxima de -13,20 %, cifras que lo posicionan a la vanguardia respecto a los enfoques alternativos evaluados. El modelo RL uniagente, aunque competitivo, obtiene resultados levemente inferiores (retorno acumulado de 2,18, Sharpe de 1,33, máx. drawdown de -18 %), evidenciando que la colaboración y coordinación multiagente permite aprovechar información dispersa, gestionar mejor la correlación entre áreas y responder de manera más eficiente ante caídas o alzas de mercado.

En contraposición, los modelos clásicos como Markowitz y Equiweight presentan un desempeño significativamente más modesto. Markowitz alcanza un retorno acumulado de 1,43, un Sharpe ratio de 0,80 y una caída máxima de -24 %, mientras que Equiweight se sitúa en 1,36, 0,77 y -22 % respectivamente. Este contraste reafirma los postulados teóricos de la moderna teoría de portafolio, la optimización dinámica y el aprendizaje adaptativo, especialmente bajo marcos multiagente, no solo maximizan el retorno esperado ajustado por riesgo, sino que logran una reducción significativa de la exposición a eventos extremos, elemento fundamental en escenarios de alta volatilidad y correlación variable entre activos.

Los resultados obtenidos validan que la inteligencia colectiva, materializada en arquitecturas MARL, constituye una evolución natural frente a los enfoques uniagente y tradicionales. El uso de agentes cooperativos permite que el sistema en su conjunto capture sinergias, minimice la probabilidad de caídas extremas y mantenga la estabilidad aun bajo condiciones de mercado adversas. Así, la evidencia empírica no solo respalda la superioridad cuantitativa del modelo MARL, sino que consolida su relevancia teórica como paradigma avanzado en la gestión de portafolios, capaz de integrar la eficiencia del retorno, la resiliencia ante el riesgo y la adaptabilidad frente a la complejidad estructural de los mercados financieros globales.

Por otro lado, el modelo Entrópico Difuso surge como un enfoque paramétrico que, a pesar de su menor sofisticación algorítmica respecto a modelos de inteligencia artificial, logra resultados competitivos, superando con claridad a los métodos clásicos en robustez y adaptabilidad. El modelo entrópico difuso implementa reglas de inferencia difusa máximizando la entropía del portafolio y las funciones de entropía difusa para el retorno y la varianza, suavizando así la toma de decisiones y atenuando los efectos de los extremos estadísticos característicos de los mercados financieros. En términos empíricos, el modelo entrópico difuso alcanza un retorno acumulado de 2.15, un Sharpe ratio de 1.24 y un drawdown máximo del 12 %, mostrando menor exposición a caídas severas que los modelos clásicos y un desempeño comparable al de los modelos RL, pero sin llegar a la eficiencia y resiliencia que logra MARL en los periodos más adversos. Su mayor aporte radica en la capacidad de construir portafolios menos frágiles y más adaptados a las preferencias reales y las carteras de los agentes , junto a sus restricciones económicas.

La integración de ambos enfoques, mediante el modelo Híbrido Entrópico Difuso, representa una superposición funcional donde se busca aprovechar la optimización adaptativa y resilencia del MARL junto con la flexibilidad ante incertidumbre del modelo difuso. Metodológicamente, el híbrido se implementa extrayendo las series de retornos periódicos generadas por los modelos RL considerándolos activos de un portafolio que se sujeta a funciones de pertenencia entrópicas difusas. Así, el modelo híbrido resulta en una serie de retornos periódicos ajustados que, al acumularse en el tiempo, exhiben un perfil competitivo al MARL con un retorno de 2.18, Sharpe de 1.33 y drawdown máximo del 15 %. Empíricamente, esta combinación logra un portafolio más robusto y alineado con las preferencias dinámicas de los agentes, especialmente en escenarios de alta volatilidad, ambigüedad o información parcial. La evidencia muestra que la curva del híbrido tiende a seguir muy de cerca al MARL en contextos estables, pero destaca en fases de alta incertidumbre, alternando posiciones de sobre-rendimiento y bajo-rendimiento respecto al MARL según la dinámica del mercado y la interacción de reglas difusas. Esta sinergia metodológica permite una gestión más equilibrada entre eficiencia cuantitativa y adaptabilidad cualitativa, integrando lo mejor de ambos mundos y abriendo la puerta a estrategias aún más resilientes y sostenibles en gestión de portafolios

Un aspecto fundamental que emerge de este capítulo es la interrogante sobre el límite y el potencial de la integración entre enfoques de inteligencia artificial distribuida y lógica difusa en la gestión de portafolios. Surge así la necesidad de explorar hasta qué punto la combinación de resiliencia multiagente y adaptabilidad difusa puede seguir induciendo mejoras en la eficiencia, estabilidad y minimización de caídas del portafolio, o si existe un umbral más allá del cual estos beneficios se estabilizan o incluso pueden enfrentarse a trade-offs imprevistos.

Esta propuesta metodológica invita a futuras investigaciones, invita a evaluar en profundidad los escenarios en que la sofisticación incremental de los modelos realmente se traduce en ganancias sustantivas y sostenibles, y en qué medida el control de la resiliencia y la incorporación explícita de la incertidumbre aportan valor en la práctica real, frente a la complejidad de los mercados modernos.

Un línea de investigación rápida podría considerarse como la integración de modelos de aprendizaje reforzado profundo multiagente con enfoques de lógica difusa, es decir considerar las salidas del modelo MARL como activos,la pregunta es: ¿los portafolios son optimizables nuevamente mediante la modelación entrópica difusa?.

Mientras el RL aporta eficiencia cuantitativa en la toma de decisiones, un esquema MARL incorpora no solo eficiencia, sino también estabilidad y resiliencia al sistema global. La inclusión de un modelo entrópico difuso sumaría la capacidad de adaptabilidad ante escenarios de incertidumbre y manejo de información parcial o ambigua, dotando al portafolio de una robustez adicional frente a eventos no anticipados, estos objetivos si bien comparten el construir un portafolio resilente, lo abordan desde objetivos y construcciones distintas. En este contexto, surge la pregunta, ¿hasta qué punto sería posible mejorar o amplificar los conceptos de resiliencia y minimización de caídas?, ¿existiría un umbral a partir del cual estos mecanismos dejarían de aportar beneficios adicionales, o su incorporación seguiría induciendo mejoras progresivas en el comportamiento del sistema? Si bien un mayor control sobre la resiliencia podría, en principio, fortalecer la protección ante caídas abruptas, también cabe investigar si existe un límite práctico más allá del cual el sistema no pueda controlar aún más el riesgo sin sacrificar otras métricas de desempeño, o si simplemente redundaría en un portafolio aún más robusto sin efectos adversos significativos. Este aspecto abre una interesante línea de investigación sobre los trade-offs entre eficiencia, resiliencia y adaptabilidad en modelos multiagente avanzados, que en esta tesis hemos llamados hídridos.

Bibliografía

[1] Harry Markowitz. "Portfolio Selection". En: *The Journal of Finance* 7.1 (1952), págs. 77-91. DOI: 10.1111/j.1540-6261.1952.tb01525.x.

[2] Stephen Boyd y Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511804441.

[3] David G. Luenberger. *Investment Science*. Oxford University Press, 1997.

[4] Rama Cont. "Empirical properties of asset returns: stylized facts and statistical issues". En: *Quantitative Finance* 1.2 (2001), págs. 223-236. DOI: 10.1080/713665670.

[5] Richard S. Sutton y Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[6] Y. Li y J. Malik. "Learning to Optimize". En: *arXiv preprint arXiv:1703.00441* (2017). arXiv: 1703.00441.

[7] Dimitri P. Bertsekas y John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[8] M. Aggarwal. "Redefining fuzzy entropy with a general framework". En: *Expert Systems with Applications* 164 (2021), pág. 113671. DOI: 10.1016/j.eswa.2020.113671.

[9] H. López-Ospina et al. "A maximum entropy optimization model for origin-destination trip matrix estimation with fuzzy entropic parameters". En: *Transportmetrica A: Transport Science* 18 (2022), págs. 963-1000. DOI: 10.1080/23249935.2021.1913257.

[10] C. B. Kalayci, O. Ertenlice y M. A. Akbay. "A comprehensive review of deterministic models and applications for mean-variance portfolio optimization". En: *Expert Systems with Applications* 125 (2019), págs. 345-368. DOI: 10.1016/j.eswa.2019.02.011.

[11] P. J. Mercurio, Y. Wu y H. Xie. "An entropy-based approach to portfolio optimization". En: *Entropy* 22 (2020), pág. 332. DOI: 10.3390/e22030332.

[12] M. Chung et al. "The effects of errors in means, variances, and correlations on the mean-variance framework". En: *Quantitative Finance* 22 (2022), págs. 1893-1903. DOI: 10.1080/14697688.2022.2083009.

[13] R. O. Michaud. "The Markowitz optimization enigma: Is 'optimized' optimal?" En: *Financial Analysts Journal* 45.1 (1989), págs. 31-42. DOI: 10.2469/faj.v45.n1.31.

[14] N. Lassance, V. DeMiguel y F. Vrins. "Optimal portfolio diversification via independent component analysis". En: *Operations Research* 70.1 (2022), págs. 55-72. DOI: 10.1287/opre.2021.2140.

[15] J. R. Yu, W. Y. Lee y W. J. P. Chiou. "Diversified portfolios with different entropy measures". En: *Applied Mathematics and Computation* 241 (2014), págs. 47-63. DOI: 10.1016/j.amc.2014.04.006.

[16] M. Cerrato et al. "Relation between higher order comoments and dependence structure of equity portfolio". En: *Journal of Empirical Finance* 40 (2017), págs. 101-120. DOI: 10.1016/j.jempfin.2016.11.007.

[17] B. Mandelbrot. "The Variation of Certain Speculative Prices". En: *The Journal of Business* 36.4 (1963), págs. 394-419. DOI: 10.1086/294632.

[18] Pankaj Gupta et al. *Multi-criteria fuzzy portfolio optimization*. Springer, 2014, págs. 161-186. DOI: 10.1007/978-3-642-54652-5_6.

[19] Y. Fang, K. K. Lai y S. Wang. *Fuzzy portfolio optimization: Theory and methods*. Vol. 609. Springer Science & Business Media, 2008. DOI: 10.1007/978-3-540-77926-1.

[20] X. Deng y X. Pan. "The research and comparison of multi-objective portfolio based on intuitionistic fuzzy optimization". En: *Computers & Industrial Engineering* 124.C (2018), págs. 411-421. DOI: 10.1016/j.cie.2018.07.044.

[21] P. Xidonas, R. Steuer y C. Hassapis. "Robust portfolio optimization: A categorized bibliographic review". En: *Annals of Operations Research* 292.1 (2020), págs. 533-552. DOI: 10.1007/s10479-020-03630-8.

[22] X. Huang. "Mean-entropy models for fuzzy portfolio selection". En: *IEEE Transactions on Fuzzy Systems* 16.4 (2008), págs. 1096-1101. DOI: 10.1109/TFUZZ.2008.924200.

[23] Z. Qin, X. Li y X. Ji. "Portfolio selection based on fuzzy cross-entropy". En: *Journal of Computational and Applied Mathematics* 228.1 (2009), págs. 139-149. DOI: 10.1016/j.cam.2008.09.010.

[24] Y. J. Liu y W. G. Zhang. "A multi-period fuzzy portfolio optimization model with minimum transaction lots". En: *European Journal of Operational Research* 242.3 (2015), págs. 933-941. DOI: 10.1016/j.ejor.2014.10.061.

[25] R. Zhou et al. "A portfolio optimization model based on information entropy and fuzzy time series". En: *Fuzzy Optimization and Decision Making* 14.4 (2015), págs. 381-397. DOI: 10.1007/s10700-015-9206-8.

[26] K. Liagkouras y K. Metaxiotis. "Multi-period mean–variance fuzzy portfolio optimization model with transaction costs". En: *Engineering Applications of Artificial Intelligence* 67 (feb. de 2018), págs. 260-269. DOI: 10.1016/j.engappai.2017.10.010.

[27] J. Zhou, X. Li y W. Pedrycz. "Mean-semi-entropy models of fuzzy portfolio selection". En: *IEEE Transactions on Fuzzy Systems* 24.6 (2016), págs. 1627-1636. DOI: 10.1109/TFUZZ.2016.2543753.

[28] B. Li y R. Zhang. "A new mean-variance-entropy model for uncertain portfolio optimization with liquidity and diversification". En: *Chaos, Solitons & Fractals* 146 (ene. de 2021), pág. 110842. DOI: 10.1016/j.chaos.2021.110842.

[29] X. Wang et al. "Multi-criteria fuzzy portfolio selection based on three-way decisions and cumulative prospect theory". En: *Applied Soft Computing* 134 (mayo de 2023), pág. 110033. DOI: 10.1016/j.asoc.2023.110033.

[30] J. Wang, H. Zhang y H. Luo. "Research on the construction of stock portfolios based on multiobjective water cycle algorithm and KMV algorithm". En: *Applied Soft Computing* 115 (ene. de 2022), pág. 108186. DOI: 10.1016/j.asoc.2021.108186.

[31] M. K. Mehlawat. "Credibilistic mean-entropy models for multi-period portfolio selection with multi-choice aspiration levels". En: *Information Sciences* 345 (jun. de 2016), págs. 9-26. DOI: 10.1016/j.ins.2016.01.079.

[32] H. Jalota, M. Thakur y G. Mittal. "A credibilistic decision support system for portfolio optimization". En: *Applied Soft Computing* 59 (oct. de 2017), págs. 512-528. DOI: 10.1016/j.asoc.2017.05.054.

[33] P. Gupta et al. "A polynomial goal programming approach for intuitionistic fuzzy portfolio optimization using entropy and higher moments". En: *Applied Soft Computing* 85 (dic. de 2019), pág. 105781. DOI: 10.1016/j.asoc.2019.105781.

[34] K. Y. Shen, H. W. Lo y G. H. Tzeng. "Interactive portfolio optimization model based on rough fundamental analysis and rational fuzzy constraints". En: *Applied Soft Computing* 125 (jul. de 2022), pág. 109158. DOI: 10.1016/j.asoc.2022.109158.

[35] R. Gómez G. "Construcción de portafolio utilizando el análisis de componentes principales y el filtro de Kalman Unscented". Tesis de maestría. Tesis de mtría. Universidad Externado de Colombia, 2019. URL: https://bdigital.uexternado.edu.co/handle/001/15806.

[36] E. Qian. "Risk parity and diversification". En: *The Journal of Investing* 20.1 (2011), págs. 119-127. DOI: 10.3905/joi.2011.20.1.119.

[37] M. López de Prado. *Machine learning for asset managers*. Cambridge: Cambridge University Press, 2019. ISBN: 9781108488084. DOI: 10.1017/9781108883658.

[38] Jie Liu, Tao Zhang y Shuyu Sun. "Review of deep learning". En: *Geoscience Frontiers* (2024).

[39] Avraam Tsantekidis et al. "Forecasting Stock Prices from the Limit Order Book Using Convolutional Neural Networks". En: *2017 IEEE 19th Conference on Business Informatics (CBI)*. Thessaloniki, Greece: IEEE, jul. de 2017, págs. 7-12. DOI: 10.1109/CBI.2017.23.

[40] Jou-Fan Chen et al. "Financial time-series data analysis using deep convolutional neural networks". En: *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*. IEEE. 2016, págs. 87-92. DOI: 10.1109/CCBD.2016.027.

[41] Wojciech Zaremba, Ilya Sutskever y Oriol Vinyals. "Recurrent Neural Networks as Adaptive Filters: Automating the Data Selection Process". En: *arXiv preprint arXiv:1409.2329* (2014). Available online at: https://arxiv.org/abs/1409.2329.

[42] Y. Macías, D. Zambrano y A. Macías. "Revisión sistemática de la literatura sobre técnicas de inteligencia artificial en la predicción de desastres naturales: Systematic review of the literature on artificial intelligence techniques in the prediction of natural disasters". En: *Revista Científica Multidisciplinar G-nerando* 5.1 (2024), págs. 1205-1235.

[43] R. Aliev, R. Abiyev y M. Menekay. "Fuzzy approach to portfolio selection using genetic algorithms". En: *Intelligent Automation & Soft Computing* 14.4 (2008), págs. 525-540.

[44] L. Di Persio y O. Honchar. "Artificial neural networks architectures for stock price prediction: Comparisons and applications". En: *International Journal of Circuits, Systems and Signal Processing* 10 (2016), págs. 403-413.

[45] Mahsa Ghorbani y Edwin K. P. Chong. "Stock price prediction using principal components". En: *PLOS ONE* 15.3 (2020), e0230124. DOI: 10.1371/journal.pone.0230124.

[46] Guangchen Wang y Zhen Wu. "The Maximum Principles for Stochastic Recursive Optimal Control Problems Under Partial Information". En: *IEEE Transactions on Automatic Control* 54.6 (jun. de 2009), págs. 1230-1242. DOI: 10.1109/TAC.2009.2019794.

[47] Ajit Kumar Rout y P. K. Dash. "Forecasting Foreign Exchange Rates Using Hybrid Functional Link RBF Neural Network and Levenberg–Marquardt Learning Algorithm". En: *Intelligent Decision Technologies* 10.3 (ago. de 2016), págs. 299-313. DOI: 10.3233/IDT-160257.

[48] Rodrigo Naranjo et al. "An Intelligent Trading System with Fuzzy Rules and Fuzzy Capital Management". En: *International Journal of Intelligent Systems* 30.8 (2015), págs. 963-983. DOI: 10.1002/int.21734.

[49] Fernando García et al. "Hybrid fuzzy neural network to predict price direction in the German DAX-30 index". En: *Technological and Economic Development of Economy* 24.6 (2018), págs. 2161-2178.

[50] Adriana Arango, Juan D Velásquez y Carlos J Franco. "Técnicas de lógica difusa en la predicción de índices de mercados de valores: una revisión de literatura." En: *Revista Ingenierías Universidad de Medellín* 12.22 (2013), págs. 117-126.

[51] John Moody y Matthew Saffell. "Learning to Trade via Direct Reinforcement". En: *IEEE Transactions on Neural Networks* 12.4 (jul. de 2001), págs. 875-889. DOI: 10.1109/72.935097.

[52] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". En: *Nature* 518.7540 (2015), págs. 529-533. DOI: 10.1038/nature14236.

[53] Zhengyao Jiang, Dixing Xu y Jinjun Liang. "A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem". En: *arXiv preprint arXiv:1706.10059* (2017). DOI: 10.48550/arXiv.1706.10059.

[54] Lucian Busoniu et al. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Automation and Control Engineering. Boca Raton, FL: CRC Press, 2010. ISBN: 978-1-4398-2108-4.

[55] Edward A Lee. "Cyber physical systems: Design challenges". En: *2008 11th IEEE international symposium on object and component-oriented real-time distributed computing (ISORC)*. IEEE. 2008, págs. 363-369.

[56] Yifu Jiang, Jose Olmo y Majed Atwi. "Deep Reinforcement Learning for Portfolio Selection". En: *Global Finance Journal* 62 (sep. de 2024), pág. 101016. DOI: 10.1016/j.gfj.2024.101016.

[57] Min Wen et al. "Stock Market Trend Prediction Using High-Order Information of Time Series". En: *IEEE Access* 7 (2019), págs. 28299-28308. DOI: 10.1109/ACCESS.2019.2901842.

[58] Maxim Lapan. *Deep Reinforcement Learning Hands-On: Apply Modern RL Methods, with Deep Q-Networks, Value Iteration, Policy Gradients, TRPO, AlphaGo Zero and More*. Birmingham, UK: Packt Publishing, 2018. ISBN: 9781788834247. DOI: 10.0000/9781788839303-001.

[59] Li Deng y Dong Yu. "Deep Learning: Methods and Applications". En: *Foundations and Trends® in Signal Processing* 7.3–4 (2014), págs. 197-387. DOI: 10.1561/2000000039.

[60] X Xiong, X Zhang J ; Jin y X Feng. "Review on financial innovations in big data era". En: *Journal of Systems Science and Information* 4.6 (2016), págs. 489-504.

[61] Zhengyao Jiang y Jinjun Liang. "Cryptocurrency Portfolio Management with Deep Reinforcement Learning". En: *2017 Intelligent Systems Conference (IntelliSys)*. London, UK: IEEE, 2017, págs. 905-913. DOI: 10.1109/IntelliSys.2017.8324237.

[62] Hee Rak Beom y Hyung Suck Cho. "A sensor-based navigation for a mobile robot using fuzzy logic and reinforcement learning". En: *IEEE transactions on Systems, Man, and Cybernetics* 25.3 (1995), págs. 464-477.

[63] Hado Van Hasselt, Arthur Guez y David Silver. "Deep reinforcement learning with double q-learning". En: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 30. 1. 2016.

[64] James M Poterba. "Stock market wealth and consumption". En: *Journal of economic perspectives* 14.2 (2000), págs. 99-118.

[65] Yi Feng, Bo Zhang y Jin Peng. "Mean-risk model for uncertain portfolio selection with background risk and realistic constraints." En: *Journal of Industrial & Management Optimization* 19.7 (2023).

[66] Yunus Ziya Kaya. "Detection of Trends and Anomalies with MACD and RSI Market Indicators for Temperature and Precipitation". En: *Symmetry* 17.8 (2025), pág. 1268.

[67] Sergey Ioffe y Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". En: *International conference on machine learning*. pmlr. 2015, págs. 448-456.

[68] Zewen Li et al. "A survey of convolutional neural networks: analysis, applications, and prospects". En: *IEEE transactions on neural networks and learning systems* 33.12 (2021), págs. 6999-7019.

[69] Jia-Hao Syu, Mu-En Wu y Jan-Ming Ho. "Portfolio management system with reinforcement learning". En: *2020 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE. 2020, págs. 4146-4151.

[70] Zhicheng Wang et al. "Deeptrader: a deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding". En: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 1. 2021, págs. 643-650.

[71] Robert E Wright. *The wealth of nations rediscovered: Integration and expansion in American financial markets, 1780-1850*. Cambridge University Press, 2002.

[72] Gordon John Alexander. *Portfolio Selection Models: a Theoretical and Empirical Investigation.* University of Michigan, 1975.

[73] Andrew J Patton. "Copula–based models for financial time series". En: *Handbook of financial time series*. Springer, 2009, págs. 767-785.

[74] Jinho Lee et al. "Global stock market prediction based on stock chart images using deep Q-network". En: *IEEE Access* 7 (2019), págs. 167260-167277.

[75] Jinho Lee et al. "Maps: Multi-agent reinforcement learning-based portfolio management system". En: *arXiv preprint arXiv:2007.05402* (2020).

[76] Lior Rokach. "A survey of clustering algorithms". En: *Data mining and knowledge discovery handbook*. Springer, 2010, págs. 269-298.

[77] Hengxi Zhang et al. "Optimizing trading strategies in quantitative markets using multi-agent reinforcement learning". En: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, págs. 136-140.

[78] Xu Wang et al. "Deep reinforcement learning: A survey". En: *IEEE Transactions on Neural Networks and Learning Systems* 35.4 (2022), págs. 5064-5078.

[79] Yuling Huang et al. "A multi-agent reinforcement learning framework for optimizing financial trading strategies based on TimesNet". En: *Expert Systems with Applications* 237 (2024), pág. 121502.

[80] Hideo Tanaka y Peijun Guo. "Portfolio selection based on upper and lower exponential possibility distributions". En: *European Journal of Operational Research* 114.1 (1999), págs. 115-126. DOI: 10.1016/S0377-2217(98)00033-2.

[81] Yanan Zheng, Ming Zhou y Gengyin Li. "Modelo de optimización difusa basado en la entropía de la información de la cartera de compras de electricidad". En: *2009 IEEE Power & Energy Society General Meeting*. Calgary, AB, Canada: IEEE, 2009, págs. 1-6. ISBN: 978-1-4244-4241-6. DOI: 10.1109/PES.2009.5275643.

[82] Kai Arulkumaran et al. "A Brief Survey of Deep Reinforcement Learning". En: *IEEE Signal Processing Magazine* 34.6 (nov. de 2017), págs. 26-38. DOI: 10.1109/MSP.2017.2743240.

[83] Christopher J. C. H. Watkins y Peter Dayan. "Q-learning". En: *Machine Learning* 8.3–4 (mayo de 1992), págs. 279-292. DOI: 10.1007/BF00992698.

[84] Richard Bellman. "Dynamic Programming". En: *Science* 153.3731 (1966), págs. 34-37. DOI: 10.1126/science.153.3731.34.

[85] Richard Stuart Sutton. "Temporal Credit Assignment in Reinforcement Learning". Order No. AAI8410337. Tesis doct. Amherst, MA, USA: University of Massachusetts Amherst, mayo de 1984, pág. 223. URL: http://www.cs.rhul.ac.uk/~chrisw/new_thesis.pdf.

[86] Christopher John Cornish Hellaby Watkins. "Learning from Delayed Rewards". Tesis doct. Cambridge, UK: King's College, University of Cambridge, mayo de 1989. URL: http://www.cs.rhul.ac.uk/~chrisw/new_thesis.pdf.

[87] Yuxi Li. "Deep reinforcement learning: An overview". En: *arXiv preprint arXiv:1701.07274* (2017).

[88] Yaodong Yang y Jun Wang. "An overview of multi-agent reinforcement learning from game theoretical perspective". En: *arXiv preprint arXiv:2011.00583* (2020).

[89] David Silver et al. "Deterministic Policy Gradient Algorithms". En: *Proceedings of the 31st International Conference on Machine Learning (ICML)*. Ed. por Eric P. Xing y Tony Jebara. Vol. 32. JMLR Workshop and Conference Proceedings. JMLR.org, jun. de 2014, págs. I–387–I-395. DOI: 10.5555/3044805.3044850.

[90] Shangding Gu et al. "A review of safe reinforcement learning: Methods, theories and applications". En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[91] Vijay R. Konda y John N. Tsitsiklis. "Actor-Critic Algorithms". En: *Advances in Neural Information Processing Systems*. Ed. por S. Solla, T. Leen y K. Müller. Vol. 12. MIT Press, 1999, págs. 1008-1014. DOI: 10.5555/3009657.3009806.

[92] Andrés García-Medina y Benito Rodríguez-Camejo. "Random Matrix Theory and Nested Clustered Optimization on High-Dimensional Portfolios". En: *International Journal of Modern Physics C* 35.08 (2024), pág. 2450098. DOI: 10.1142/S0129183124500980.

[93] Andrés García-Medina, Salvatore Micciché y Rosario N. Mantegna. "Two-step Estimators of High-Dimensional Correlation Matrices". En: *Physical Review E* 108.4 (2023), pág. 044137. DOI: 10.1103/PhysRevE.108.044137.

[94] A. García-Medina y R. Macías Páez. "Rotationally invariant estimators on portfolio optimization to unveil financial risk's states". En: *International Journal of Modern Physics C* 34.09 (2023), págs. 1-19. DOI: 10.1142/S0129183123501176.

[95] Atsushi Inoue, Lu Jin y Barbara Rossi. "Rolling Window Selection for Out-of-Sample Forecasting with Time-Varying Parameters". En: *Journal of Econometrics* 196.1 (2017), págs. 55-67. DOI: 10.1016/j.jeconom.2016.03.006.

[96] Yuanyuan Zhang, Xiang Li y Sini Guo. "Portfolio Selection Problems with Markowitz's Mean–Variance Framework: A Review of Literature". En: *Fuzzy Optimization and Decision Making* 17.2 (jun. de 2018), págs. 125-158. DOI: 10.1007/s10700-017-9266-z.

[97] Kaisa Miettinen. *Nonlinear Multiobjective Optimization*. 1.ª ed. Vol. 12. International Series in Operations Research & Management Science. New York, NY: Springer, 1998. ISBN: 978-0-7923-8278-2. DOI: 10.1007/978-1-4615-5563-6.

[98] Drew Fudenberg y Jean Tirole. *Game Theory*. Cambridge, MA: MIT Press, 1991. ISBN: 9780262061414.

[99] Sujoy Dhar. "Bond Market: Untapped Potential for Investors". En: (2008).

[100] Robin Greenwood y David Thesmar. "Stock price fragility". En: *Journal of Financial Economics* 102.3 (2011), págs. 471-490.

[101] Robin Greenwood y David Scharfstein. "The growth of finance". En: *Journal of Economic perspectives* 27.2 (2013), págs. 3-28.

[102] Lidia Bolla, Alexander Kohler y Hagen Wittig. "Index-linked investing–A curse for the stability of financial markets around the globe?" En: *Journal of Portfolio Management* 42.3 (2016), pág. 26.

[103] Marcel Kahan y Edward B Rock. "Index funds and corporate governance: Let shareholders be shareholders". En: *BuL rev.* 100 (2020), pág. 1771.

[104] Hong Liu y Yajun Wang. *Index investing and price discovery*. SSRN, 2018.

[105] Ilia D Dichev y Gwen Yu. "Higher risk, lower returns: What hedge fund investors really earn". En: *Journal of Financial Economics* 100.2 (2011), págs. 248-263.

[106] Bilal Hafeez, M Humayun Kabir y Udomsak Wongchoti. "Are retail investors really passive? Shareholder activism in the digital age". En: *Journal of Business Finance & Accounting* 49.3-4 (2022), págs. 423-460.

[107] Pawel Fiedor y Petros Katsoulis. *Information and liquidity linkages in EFTs and underlying markets*. Inf. téc. Central Bank of Ireland, 2020.

[108] Luca J Liebi. "The effect of ETFs on financial markets: a literature review". En: *Financial Markets and Portfolio Management* 34.2 (2020), págs. 165-178.

[109] Chongjie Zhang y Victor Lesser. "Coordinating multi-agent reinforcement learning with limited communication". En: *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. 2013, págs. 1101-1108.

[110] Xiao-Yang Liu et al. "FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance". En: *arXiv preprint arXiv:2011.09607* (2020).

[111] Ziyuan Zhou, Guanjun Liu y Ying Tang. "Multi-agent reinforcement learning: Methods, applications, visionary prospects, and challenges". En: *arXiv preprint arXiv:2305.10091* (2023).

[112] Xiao-Yang Liu et al. "Dynamic datasets and market environments for financial reinforcement learning". En: *Machine Learning* 113.5 (2024), págs. 2795-2839.

[113] John Christopher Tidwell y John Storm Tidwell. "Deep Q-Network (DQN) multi-agent reinforcement learning (MARL) for Stock Trading". En: *arXiv preprint arXiv:2505.03949* (2025).

[114] Laura Del Amo Albiol. "An Ensemble of Deep Reinforcement Learning algorithms for trading Exchange-Traded Funds". En: (2024).

[115] Frank J. Fabozzi et al. *Robust Portfolio Optimization and Management*. Hoboken, NJ: John Wiley & Sons, 2007. ISBN: 9780471921226.

[116] Xi Chen y Xiaotie Deng. "Settling the Complexity of Two-Player Nash Equilibrium". En: *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006)*. Berkeley, CA, USA: IEEE, oct. de 2006, págs. 261-272. ISBN: 0-7695-2720-5. DOI: 10.1109/FOCS.2006.69.