Protocolo para la Presentación del Anteproyecto de Grado Opción <u>Monografía</u>

* Este espacio lo diligencia el área de trabajos de grado
Referencia del proyecto* _____

Información General

Información del estudiante 1	Nombre: Luis Fernando Mercado
	Programa de maestría: Ciencia de Datos
	Tipo y número de documento: CC. 1107053810
	Correo institucional: lmercad53810@universidadean.edu.co
	Teléfonos de contacto: 3016730684
	Correo electrónico alternativo: luisferemprende@gmail.com
Información del estudiante 2	Nombre:
	Programa de maestría:
	Tipo y número de documento:
	Correo institucional:
	Teléfonos de contacto:
	Correo electrónico alternativo:
Información del estudiante 3	Nombre:
	Programa de maestría:
	Tipo y número de documento:
	Correo institucional:
	Teléfonos de contacto:
	Correo electrónico alternativo:
Campo de investigación[1]:	Machine Learning
Grupo de investigación[1]:	Finanzas
Línea de investigación[1]:	Deep Learning
Título tentativo del proyecto:	*Estrategia de ensamble con aprendizaje por refuerzo profundo para la gestión de portafolios en el mercado accionario colombiano: selección dinámica entre PPO, A2C y DDPG mediante la razón de Sharpe en ventanas móviles*

[1] https://universidadean.edu.co/investigacion/grupos-de-investigacion

Lugar y fecha de presentación Bogota, 20 de Mayo de 2026

Tabla de contenido

Planteamiento del Problema 3
Objetivos 5
Marco Institucional 7
Marco teórico 10
Hipótesis 15
Variables 15
Diseño metodológico 17
Resultados preliminares de investigación 28
Conclusiones Preliminares 32
Recomendaciones preliminares 35
Cronograma 37
Referencias 40

Planteamiento del Problema

- **Antecedentes investigativos y contexto.**

Los mercados financieros presentan dinámicas no estacionarias en las que la volatilidad, las correlaciones y la microestructura cambian de forma persistente, generando regímenes que afectan el desempeño de cualquier estrategia sistemática de inversión. En este contexto, un único agente de aprendizaje por refuerzo profundo puede mostrar resultados sobresalientes en un periodo determinado y deteriorarse de manera importante en otro, debido a que la política aprendida puede sobreajustarse a patrones transitorios del mercado sobre el que fue entrenado.

Esta barrera ha motivado, de acuerdo con la literatura reciente, el desarrollo de estrategias más flexibles que buscan combinar agentes con comportamientos distintos para responder mejor a entornos dinámicos, no estacionarios y cambiantes. Un antecedente relevante es el de Yang, Liu, Zhong y Walid (2020) y Jiang, Xu y Liang (2017), quienes propusieron una estrategia de ensamble con PPO, A2C y DDPG y encontraron que dicha combinación podía superar a los agentes individuales y a estrategias de referencia en términos de desempeño ajustado por riesgo medido con la razón de Sharpe, evaluando sobre acciones del índice Dow Jones.

Por su parte, Jiang, Xu y Liang (2017) propusieron un marco de DRL para la gestión de portafolios financieros, mostrando que distintas arquitecturas podían superar estrategias comparadas en ejercicios de backtesting. Lee, Kim, Yi y Kang (2020) desarrollan un sistema multiagente de gestión de portafolios que tiene en cuenta principios de diversificación, mientras que Xu et al. (2020) introducen una arquitectura relation-aware transformer para capturar de mejor manera los patrones secuenciales y las correlaciones entre activos. En literatura reciente, Jiang, Olmo y Atwi (2024) plantean un marco de DRL para selección de

portafolios en escenarios de alta dimensión, incluyendo la aversión al riesgo y costos de transacción en la función de recompensa.

En el contexto colombiano, Mejía Estrada (2023) aplicó la técnica de aprendizaje por refuerzo profundo a portafolios de renta fija, mientras que Obando Morales (2022) y Quintero González (2024) exploraron aplicaciones en criptoactivos y pair-trading, evidenciando así, un interés académico creciente, aunque todavía sin una evaluación específica para portafolios accionarios locales.

- **Descripción del problema.**

A pesar de los avances mencionados, el caso colombiano presenta condiciones particulares que dificultan la transferencia directa de metodologías desarrolladas en mercados de alta liquidez. Entre estas dificultades, confluyen al menos tres factores, el primero de ellos es la disponibilidad limitada de datos históricos consistentes para algunos activos, la segunda, es la menor liquidez relativa y los spreads más amplios, y, la tercera, alta sensibilidad del mercado ante episodios de estrés financiero, cambios estructurales e incertidumbre económica. Estas condiciones elevan el costo efectivo de las operaciones, incrementan el riesgo de sobreajuste y limitan la capacidad de generalización de los modelos.

Frente a esta situación, la alternativa metodológica consiste en emplear un ensamble de agentes de aprendizaje por refuerzo profundo con capacidades complementarias, en este caso particular, PPO, A2C y DDPG, y seleccionar dinámicamente el agente activo con base en su desempeño reciente fuera de muestra, medido mediante la razón de Sharpe en ventanas móviles de validación. Esta estrategia permite que la decisión no dependa de un único modelo durante todo el horizonte temporal, sino que, puede adaptarse a cambios en las condiciones del mercado (Xu et al. 2020; Jiang, Olmo y Atwi, 2024). Además, la incorporación explícita de costos de transacción y de un mecanismo de control de riesgo

por turbulencia o régimen hace posible evaluar el desempeño del ensamble bajo condiciones más cercanas a la operación real de un portafolio accionario colombiano.

- **Delimitación del problema.**

De acuerdo con lo anterior, la investigación se delimita al mercado accionario colombiano y la gestión de portafolios a partir de series históricas de activos con criterios de continuidad y liquidez. No se abordarán todos los instrumentos del mercado de valores ni estrategias intradía o de alta frecuencia, sino un problema de asignación de pesos en portafolio con datos históricos y evaluación fuera de muestra mediante backtesting – como se explicará más adelante-. En esa misma línea, el estudio se centrará en comparar el desempeño del ensamble frente a los agentes individuales y a estrategias de referencia, considerando como variables centrales el retorno acumulado, la razón de Sharpe, la volatilidad, el máximo drawdown, los costos de transacción y el control de riesgo por turbulencia. Esta delimitación permite acotar el alcance del proyecto y hacerlo metodológicamente viable, al tiempo que conserva pertinencia para el análisis de decisiones de inversión en un contexto bursátil emergente.

- **Pregunta de investigación.**

¿En qué medida un ensamble de aprendizaje por refuerzo profundo para la gestión de portafolios en el mercado accionario colombiano, con selección dinámica entre PPO, A2C y DDPG mediante la razón de Sharpe en ventanas móviles, mejora el desempeño ajustado por riesgo bajo costos de transacción y control de riesgo por turbulencia o régimen?

Objetivos

Objetivo general.

Evaluar el desempeño ajustado por riesgo de un ensamble de aprendizaje por refuerzo profundo para la gestión de portafolios en el mercado accionario colombiano, con selección

dinámica entre PPO, A2C y DDPG mediante la razón de Sharpe en ventanas móviles, incorporando costos de transacción y control de riesgo por turbulencia.

Objetivos específicos.

- Construir un conjunto de datos del mercado accionario colombiano con criterios de continuidad, liquidez y tratamiento de datos faltantes.
- Implementar un entorno de gestión de portafolios para el entrenamiento de los agentes PPO, A2C y DDPG bajo condiciones homogéneas de simulación.
- Contrastar el desempeño del ensamble frente a los agentes individuales y a estrategias de referencia mediante métricas de retorno y riesgo bajo condiciones cambiantes del mercado colombiano.

Justificación

Desde el punto de vista teórico, la aplicación del aprendizaje por refuerzo profundo a la gestión de portafolios constituye una de las líneas de investigación de mayor dinamismo en la intersección entre inteligencia artificial y finanzas cuantitativas. Sin embargo, gran parte de la evidencia disponible se concentra en mercados de alta liquidez, como el estadounidense o el chino, caracterizados por una mayor disponibilidad de datos y menores fricciones operativas. En ese sentido, trasladar y validar este enfoque en el mercado colombiano representa un aporte relevante para comprender los límites de generalización de los modelos DRL en contextos con mayores restricciones de liquidez, costos de transacción más altos y series de datos más escasas.

Metodológicamente, el estudio contribuye a fortalecer los criterios de evaluación de estrategias basadas en DRL en el ámbito académico nacional. La incorporación explícita de costos de transacción, la selección dinámica del agente activo con base en su desempeño fuera de muestra y el uso de mecanismos de control de riesgo responden a

cuestionamientos recurrentes sobre el sesgo retrospectivo y el sobreajuste en ejercicios de backtesting. En consecuencia, el diseño propuesto busca establecer un protocolo reproducible que pueda servir como referencia para futuras investigaciones en esta línea.

Los resultados de esta investigación pueden orientar a gestores de portafolios, firmas comisionistas y fondos de inversión colombianos en la adopción de herramientas de inteligencia artificial para la toma de decisiones de inversión, con criterios claros sobre las condiciones de mercado bajo las cuales dichas herramientas agregan valor. De esta manera, la evidencia sobre la sensibilidad a costos y a cambios de régimen puede aportar criterios útiles para la gestión del riesgo operativo de estrategias cuantitativas en el entorno local.

Viabilidad del proyecto

Tabla 1. Criterios de factibilidad del proyecto.

Criterio	***Factibilidad (siendo 1 menor y 5 mayor)***
Acceso a la información	5
Apoyo e interés de colaboradores o aliados	5
Disponibilidad de recursos requeridos	5
Probabilidad de avance en el tiempo establecido	5
Probabilidad de continuidad o implementación de la propuesta de investigación	5
Promedio	5

Marco Institucional

El contexto bajo el cual se desarrollará la presente investigación es el mercado de valores colombiano. El cual, a su vez tiene como marco general de funcionamiento la Ley 964 de 2005, la cual establece tanto los objetivos como los criterios de intervención del

Estado sobre las actividades de aprovechamiento, manejo e inversión de recursos captados del público a través de valores (Congreso de la República de Colombia, 2005). Esta ley orienta también la regulación del mercado hacia la protección de los inversionistas, así como la preservación de la transparencia, la integridad y la confianza. Así mismo, el Decreto 2555 de 2010 reúne en una sola norma las disposiciones aplicables al sector financiero, asegurador y al mercado de valores, por lo que constituye la principal referencia regulatoria para el análisis institucional de esta propuesta (Presidencia de la República de Colombia, 2010).

En adición a la normativa mencionada, la Superintendencia Financiera de Colombia de forma institucional, tiene el rol de autoridad de inspección, vigilancia y control. De acuerdo con la Ley 964 de 2005, debe instruir a las entidades sujetas a supervisión sobre la forma en que deben, por un lado, cumplir las disposiciones del mercado de valores, y por otro, fijar criterios técnicos y jurídicos para aplicación y adopción de medidas preventivas asociadas al riesgo que pueda haber para los inversionistas o para la integridad del mercado. De igual forma, el Sistema Integral de Información del Mercado de Valores - SIMEV, es el encargado de articular los registros relacionados, entre ellos, el Registro Nacional de Valores y Emisores – RNVE, el Registro Nacional de Agentes del Mercado de Valores y el Registro Nacional de Profesionales del Mercado de Valores, los cuales, ordenan la participación de los emisores, los intermediarios y los profesionales habilitados.

El mercado colombiano tiene instancias de autorregulación y organización como la Bolsa de Valores de Colombia – BVC y el Autorregulador del Mercado de Valores de Colombia – AMV. La primera concentra una parte esencial de la infraestructura del mercado, ya que, es la encargada de publicar la compilación de leyes, decretos y circulares que rigen el funcionamiento de dicho mercado, incluyendo la Ley 964 de 2005 ya mencionada (Bolsa de Valores de Colombia - BVC, s.f.). Por su parte, el segundo, es una organización de

carácter privado, sin ánimo de lucro, que se encarga regular, monitorear, disciplinar y profesionalizar el mercado de valores colombiano.

Ahora, haciendo más énfasis en la normatividad asociada a la gestión de portafolios, es relevante mencionar la regulación sobre administración de portafolios de terceros. En ese sentido, el Decreto 1247 de 2016 define esta actividad como aquella mediante la cual "las sociedades comisionistas de bolsa reciben activos de un tercero con la finalidad de administrar un portafolio a su criterio, pero con base en los lineamientos y objetivos dispuestos por dicho tercero" (Presidencia de la República de Colombia, 2016). Es decir, que son receptores de activos de un tercero para administrar un portafolio a su criterio, pero conforme a los lineamientos y objetivos fijados por el cliente. Este mismo Decreto señala que esta actividad requiere la previa autorización de la Superintendencia Financiera de Colombia y, además, exige que la entidad actúe con la diligencia propia de un experto.

En complemento, la regulación más reciente ha seguido ajustando el entorno del mercado de valores colombiano. Por ejemplo, el Decreto 661 de 2018 precisa que la actividad de asesoría hace parte del mercado de valores y, además resalta la necesidad de adecuar las obligaciones de las entidades vigiladas a la evolución de estos instrumentos (Presidencia de la República de Colombia, 2018). Más actuales, los Decretos 265 y 1239 de 2024 introdujeron modificaciones al Decreto 2555 de 2010 en asuntos relacionados, entre otros, con los fondos de inversión colectiva, los sistemas de cotización de valores extranjeros y los instrumentos que contribuyen a la liquidez del mercado (Presidencia de la República de Colombia, 2024a; 2024b). Estas actualizaciones no pueden ser soslayadas, ya que, muestran que el análisis de estrategias cuantitativas no puede separarse del desarrollo institucional del mercado en el que se pretende evaluar.

Marco teórico

Hay algunos enfoques teóricos y conceptuales que son necesarios desarrollar para orientar no solo un marco de investigación sino también sustentar y orientar el diseño metodológico. En primer lugar, se abordará la gestión de portafolios como un problema de decisión bajo riesgo e incertidumbre, esto se hace con el fin de poder establecer la base financiera desde la cual se evalúa la relación entre retorno y riesgo. En segundo lugar, como ya se ha mencionado, se incorporará el aprendizaje por refuerzo como marco necesario para modelar las decisiones secuenciales en entornos dinámicos, puesto que uno de los propósitos de la presente investigación es representar la adaptación del portafolio a cambios en el mercado. En tercer lugar, presentan los métodos actor-crítico, particularmente PPO, A2C y DDPG, debido a que constituyen la base algorítmica del estudio. Posteriormente, la lógica del ensamble, las ventanas móviles y la evaluación fuera de muestra, dado que estas permitirán justificar la selección dinámica entre agentes en contextos no estacionarios. Finalmente, se exponen tanto el desempeño ajustado por riesgo, como los costos de transacción y el control de riesgo, ya que, son elementos propicios para aproximar la evaluación del modelo a condiciones más cercanas a la realidad del mercado accionario colombiano.

Gestión de portafolios como problema de decisión bajo riesgo e incertidumbre.

La selección de portafolios no puede entenderse únicamente como la búsqueda del mayor retorno esperado, sino como una decisión que exige ponderar conjuntamente rentabilidad y riesgo. En este sentido, Markowitz (1952) es quien formaliza la asignación de activos como un problema de elección entre combinaciones posibles de retorno y varianza, dejando así, el precedente para la evaluación de la conveniencia de distintas composiciones del portafolio. Si bien, entre los propósitos del documento no está desarrollar desde la optimización media-varianza como comúnmente se hace, sí es necesario traer a colación

que la calidad de una estrategia de inversión va a depender de su capacidad para manejar el equilibrio entre la ganancia esperada y la exposición al riesgo.

Aprendizaje por refuerzo profundo como marco para decisiones secuenciales de portafolio.

El aprendizaje por refuerzo permite la modelación de la gestión de portafolios como un problema de decisión secuencial en el que un agente al interactuar con un entorno observa estados, ejecuta acciones y aprende a maximizar una recompensa acumulada a lo largo del tiempo (Sutton & Barto, 2018). Esta formulación permite representar de una forma más integral las variables del mercado, decisiones de rebalanceo y los resultados del portafolio dentro de un mismo entorno de aprendizaje, lo cual es útil para la evaluación de estrategias adaptativas bajo incertidumbre (Liu et al., 2020).

Además, porque desarrollos aplicados como FinRL muestran que este tipo de entornos puede agregar puntualmente restricciones relevantes para la metodología del estudio, entre ellas los costos de transacción, la liquidez del mercado y el grado de aversión al riesgo del inversionista (Liu et al., 2020). Por ello, en esta investigación el aprendizaje por refuerzo no se propone solo como una técnica computacional, sino como un enfoque teórico/metodológico que permite sustentar el entrenamiento, la comparación y la selección dinámica de agentes en la gestión de portafolios del mercado accionario colombiano.

Algoritmos Proximal Policy Optimization - PPO, Advantage Actor-Critic - A2C y Deep Deterministic Policy Gradient - DDPG.

Los tres algoritmos empleados en esta investigación representan enfoques complementarios dentro del DRL. El algoritmo PPO es un método de política de gradiente que actualiza la política del agente de forma conservadora mediante una función objetivo recortada, lo que lo hace estable y adecuado para entornos con espacios de acción continuos y recompensas ruidosas (Schulman et al., 2017). El algoritmo A2C combina una

función de valor (critic) con una política (actor) y utiliza múltiples entornos en paralelo para reducir la varianza de las estimaciones del gradiente (Mnih et al., 2016). Por su parte, el algoritmo DDPG es un método de actor-crítico diseñado específicamente para espacios de acción continuos, que emplea redes neuronales profundas y una memoria de repetición de experiencias para lograr un aprendizaje estable (Lillicrap et al., 2016).

Métodos de actor-crítico.

La selección de PPO, A2C y DDPG son algoritmos que pertenecen a la familia de métodos actor-crítico, bajo la cual la estructura de aprendizaje estima la política de decisión y otra valora los estados o acciones para orientar su actualización (Mnih et al., 2016; Lillicrap et al., 2016). Esta arquitectura es pertinente para los fines de esta investigación, ya que, la gestión de portafolios exige decisiones adaptativas en entornos dinámicos, más allá de reglas fijas de asignación de activos. En esa misma línea, A2C constituye un antecedente importante al derivarse de los métodos actor-crítico asincrónicos, los cuales son propuestos por Mnih et al. (2016), en donde la contribución fue mostrar que múltiples procesos de aprendizaje pueden estabilizar el entrenamiento y mejorar el desempeño del agente.

Por su parte, PPO introduce una función objetivo que limita los cambios bruscos en la política y favorece actualizaciones más estables durante el entrenamiento (Schulman, Wolski, Radford, & Klimov, 2017). Mientras que, DDPG cuando las decisiones se representan en espacios de acción continuos, es una alternativa que combina el enfoque actor-crítico con gradientes determinísticos de política para aprender controles de manera eficiente (Lillicrap et al., 2016).

Ensamble, ventanas móviles y evaluación fuera de muestra.

Se debe tener una premisa conceptual en contextos financieros no estacionarios para la presente investigación, y es que no existe garantía de que un solo agente pueda mantener un desempeño superior en todos los periodos. A partir de dicha premisa, el uso de un

ensamble permite comparar agentes con comportamientos distintos y seleccionar de manera periódica aquel que muestre mejores resultados recientes bajo un criterio homogéneo de evaluación, lo cual evidencia su pertinencia y aplicabilidad para estos casos.

Yang et al. (2020) propusieron una estrategia de ensamble entre PPO, A2C y DDPG para selección dinámica en inversión automatizada, mostrando que la combinación de estos algoritmos puede estar por encima tanto de los agentes individuales como de las estrategias de referencia en términos de desempeño ajustado por riesgo. En consecuencia, la evaluación mediante ventanas móviles de entrenamiento, validación y prueba toma relevancia teórica y metodológica, pues permite valorar el comportamiento de los agentes en distintos momentos del mercado y reduce el riesgo de extraer conclusiones a partir de un único periodo favorable.

Desempeño ajustado por riesgo, costos de transacción y control de riesgo.

La evaluación del ensamble requiere una medida que no solo observe el retorno obtenido, sino también la estabilidad relativa que se obtendrá con dicho retorno. Por eso, la razón de Sharpe constituye un referente apropiado, debido a que, expresa el exceso de retorno por unidad de variabilidad y permite comparar estrategias desde una perspectiva ajustada por riesgo (Sharpe, 1994). Esto es coherente con la investigación, puesto que, la selección dinámica entre agentes no busca identificar solo cuál gana más en determinado periodo, sino cómo es su comportamiento al asumir un riesgo, no obstante, una evaluación muy realista del desempeño no puede desligarse de las fricciones del mercado.

En ese sentido, la literatura que habla sobre el aprendizaje por refuerzo aplicado a finanzas ha mostrado que añadir costos de transacción modifica el resultado de las estrategias, al desincentivar rotaciones excesivas y aproximar la simulación a condiciones de operación más manejables (Deng, Feng, Kong, Ren, & Dai, 2017). Por otro lado, en cuanto a los mecanismos de control de riesgos, Kritzman y Li (2010) proponen el concepto

de turbulencia financiera, que permite identificar episodios anómalos del comportamiento de los activos y ajustar la exposición del portafolio cuando el entorno cambia de manera abrupta.

Mercados emergentes, liquidez y fricciones de negociación.

El análisis de estrategias cuantitativas de inversión no puede soslayarse de las características estructurales del mercado en el que se aplican. Por ejemplo, en mercados emergentes, la menor liquidez relativa, la concentración de la negociación en un número reducido de activos, la presencia de spreads más amplios y la mayor sensibilidad a choques macroeconómicos o políticos hacen que la operación real de un portafolio enfrente barreras diferentes a las observadas en mercados más profundos y desarrollados. Aquí es importante esta distinción, ya que una estrategia que obtiene resultados favorables en entornos de alta liquidez puede perder robustez cuando se traslada a mercados donde las fricciones y/o barreras operativas tienen un peso mayor sobre el retorno efectivo. En ese sentido, la literatura reafirma la importancia de incorporar condiciones de mercado más realistas dentro del entorno de simulación, especialmente cuando se busca evitar que el agente aprenda reglas de decisión artificialmente rentables bajo supuestos de negociación sin fricciones (Liu et al., 2020; Deng et al., 2017).

Para los fines de esta investigación, entender estas condiciones de antemano es importante porque el mercado accionario colombiano presenta un contexto diferente al de los escenarios en los que se han desarrollado buena parte de los estudios internacionales de referencia. De esta manera, la evaluación del ensamble no se trata solamente de centrarse únicamente en su capacidad de generar retorno, sino también en su comportamiento bajo restricciones asociadas a liquidez, costos de transacción y episodios de inestabilidad. De esta forma, el análisis no solo se apoya en la lógica general del aprendizaje por refuerzo profundo, sino en una lectura contextualizada de su aplicación a

un mercado emergente como el colombiano, donde la robustez operativa del modelo es tan importante como su desempeño ajustado por riesgo (Yang et al., 2020; Liu et al., 2020).

Hipótesis

Hipótesis general.

Un ensamble de aprendizaje por refuerzo profundo para la gestión de portafolios en el mercado accionario colombiano, con selección dinámica entre PPO, A2C y DDPG mediante la razón de Sharpe en ventanas móviles, presenta un mejor desempeño ajustado por riesgo que los agentes individuales y las estrategias de referencia, incorporando costos de transacción y control de riesgo por turbulencia.

Hipótesis complementaria.

La incorporación de costos de transacción y de un mecanismo de control de riesgo por turbulencia permite que el ensamble mantenga un comportamiento más robusto y estable fuera de muestra, reflejado en mejores niveles de razón de Sharpe y menores valores de volatilidad y drawdown frente a esquemas que no tienen estos ajustes.

Variables

Variable	Definición conceptual	Operacionalización	Clasificación
Tipo de estrategia de portafolio	Modalidad de decisión utilizada para asignar pesos a los activos del portafolio. Permite distinguir entre el ensamble DRL, los agentes individuales y las estrategias de referencia.	Se registrará para cada ventana de prueba la estrategia evaluada: ensamble DRL, PPO, A2C, DDPG, COLCAP y estrategia de igual ponderación.	Independiente; cualitativa nominal politómica

Variable	Definición conceptual	Operacionalización	Clasificación
Desempeño ajustado por riesgo	Relación entre la rentabilidad obtenida y el nivel de riesgo asumido por la estrategia. Constituye la variable principal de evaluación del estudio.	Se medirá principalmente mediante la razón de Sharpe calculada por ventana y para el período agregado, utilizando tasa libre de riesgo anual de referencia.	Dependiente principal; cuantitativa continua
Retorno acumulado	Ganancia o pérdida total generada por una estrategia durante el período de prueba.	Se calculará como el crecimiento porcentual acumulado del valor del portafolio en cada ventana y en el período total fuera de muestra.	Dependiente complementaria; cuantitativa continua
Volatilidad anualizada	Medida de dispersión de los rendimientos del portafolio, utilizada como aproximación del riesgo total.	Se estimará a partir de la desviación estándar de los rendimientos diarios anualizados de cada estrategia.	Dependiente complementaria; cuantitativa continua
Máximo drawdown	Máxima caída acumulada del valor del portafolio desde un pico previo hasta un valle posterior dentro del período analizado.	Se calculará para cada estrategia y ventana como la mayor pérdida relativa observada durante el backtesting.	Dependiente complementaria; cuantitativa continua
Costos de transacción	Fricción operativa derivada del rebalanceo del portafolio, asociada a comisiones y spreads de negociación.	Se modelarán como un costo proporcional al cambio en los pesos del portafolio en cada operación. Su nivel base se fijará en la simulación y podrá variarse en el análisis de sensibilidad.	Variable de control; cuantitativa continua
Turbulencia financiera	Medida del grado de anomalía estadística del mercado en un momento dado, útil para identificar episodios de estrés.	Se calculará mediante el índice de turbulencia financiera con base en la distancia de Mahalanobis sobre retornos históricos. Cuando supere el umbral definido, activará una regla defensiva en el portafolio.	Variable de control/contextual; cuantitativa continua

Variable	Definición conceptual	Operacionalización	Clasificación
Rotación del portafolio (turnover)	Intensidad con la que una estrategia modifica la composición del portafolio a lo largo del tiempo.	Se estimará a partir del cambio absoluto en los pesos del portafolio entre períodos consecutivos. Servirá para interpretar el impacto operativo de los costos de transacción.	Variable descriptiva complementaria; cuantitativa continua

Diseño metodológico

La presente investigación se desarrolla desde un enfoque cuantitativo y aplicado, orientado a evaluar si un ensamble de aprendizaje por refuerzo profundo mejora el desempeño ajustado por riesgo en la gestión de portafolios del mercado accionario colombiano. Para ello, se construirá una base de datos a partir de activos seleccionados con criterios de continuidad de la serie, liquidez relativa y disponibilidad de información histórica. Sobre esta base, la gestión de portafolios se modelará como un problema de decisión secuencial en el que un agente observa un estado del entorno, ejecuta una acción de asignación y recibe una recompensa asociada al comportamiento del portafolio, siguiendo la lógica del aprendizaje por refuerzo y su aplicación a entornos financieros reproducibles (Sutton & Barto, 2018; Liu et al., 2020).

Sobre el entorno definido se entrenarán tres agentes de aprendizaje por refuerzo profundo: Advantage Actor-Critic (A2C), Proximal Policy Optimization (PPO) y Deep Deterministic Policy Gradient (DDPG). La elección de estos algoritmos responde a que pertenecen a la familia de métodos actor-crítico y ofrecen comportamientos diferenciados para el aprendizaje de políticas en contextos dinámicos. En términos metodológicos, A2C aporta una referencia estable dentro de esta lógica; PPO resulta pertinente por su mecanismo de optimización de políticas mediante una función objetivo que restringe actualizaciones excesivas; y DDPG es especialmente útil cuando las decisiones se

representan en espacios de acción continuos, como ocurre con la asignación de pesos dentro de un portafolio (Mnih et al., 2016; Schulman et al., 2017; Lillicrap et al., 2016).

La comparación entre los agentes se realizará mediante un esquema de ventanas móviles de entrenamiento, validación y prueba. En cada iteración, los tres modelos serán entrenados sobre una ventana histórica, evaluados en una ventana posterior de validación y, con base en ese resultado, se seleccionará el agente activo para el siguiente periodo de prueba. El criterio principal de selección será la razón de Sharpe, dado que permite comparar el desempeño ajustado por riesgo y no únicamente el retorno bruto obtenido en cada periodo (Sharpe, 1994). Esta lógica retoma la propuesta de ensamble desarrollada por Yang et al. (2020), quienes muestran que la combinación dinámica de PPO, A2C y DDPG puede ofrecer un mejor desempeño ajustado por riesgo que el uso aislado de cualquiera de estos algoritmos.

Con el propósito de aproximar la evaluación a condiciones más realistas del mercado colombiano, la simulación incorporará costos de transacción explícitos y un mecanismo de control de riesgo por turbulencia o régimen. La inclusión de costos busca penalizar estrategias con rotación excesiva y evitar que los resultados dependan de supuestos irreales de fricción nula, mientras que el control por turbulencia permitirá reducir exposición en episodios de alta inestabilidad del mercado (Deng et al., 2017; Kritzman & Li, 2010). Finalmente, el desempeño del ensamble será comparado con el de los agentes individuales y con estrategias de referencia mediante métricas como retorno acumulado, volatilidad anualizada y máximo drawdown, complementadas por la razón de Sharpe como indicador central de evaluación. De manera adicional, se realizará un análisis de sensibilidad sobre el tamaño de las ventanas, los niveles de costos de transacción y las reglas de control de riesgo, con el fin de examinar la robustez del diseño metodológico propuesto.

Población, muestra y fuentes de información.

Se utilizarán tres fuentes de datos de precios de cierre para las acciones que integran el índice COLCAP. En primer lugar, se emplearon archivos CSV históricos descargados directamente desde Investing.com para Ecopetrol (EC), PFBCOLOM, Grupo Argos y Grupo Sura Pref, dado que estos cuatro tickers presentaban inconsistencias en las fuentes automáticas. En segundo lugar, se utilizó Stooq como fuente secundaria para el resto de activos, accediendo mediante el sufijo .co que identifica acciones de la Bolsa de Valores de Colombia. En tercer lugar, Yahoo Finance sirvió como fuente de respaldo cuando Stooq no alcanzaba el umbral mínimo de cobertura del 50 % del período de análisis. El benchmark del índice COLCAP se descargó igualmente desde Investing.com en formato CSV. El período de análisis comprende desde enero de 2015 hasta diciembre de 2025, lo que equivale a 10 años de datos bursátiles colombianos y 29 ventanas de validación y prueba en el esquema móvil adoptado.

Técnicas de recolección de datos.

La recolección de datos se hará acorde al uso de series históricas del mercado accionario colombiano y con la lógica de simulación planteada en el diseño metodológico. En primer lugar, se utilizará la revisión documental de fuentes secundarias, mediante la consulta y extracción de información histórica de mercado, entre ella precios de cierre, precios ajustados, volúmenes negociados y otras variables que se encuentren necesarias para representar el comportamiento de los activos seleccionados, construyendo así, la base de datos inicial del estudio. Por ejemplo, Yang et al. (2020) a partir de 30 acciones del Dow Jones construyen la estrategia de ensamble, mientras que Jiang, Olmo y Atwi (2024) lo hacen a partir de componentes del Dow Jones Industrial Average y del S&P100.

En segundo lugar, se aplicará una técnica de registro estructurado de series temporales, con el fin de organizar la información recolectada de acuerdo con criterios homogéneos de fecha, activo y variable. Una vez recolectados los datos brutos, se aplicará un protocolo de

homogeneización y depuración coherente con las recomendaciones metodológicas de la literatura sobre DRL en finanzas (Liu et al., 2020; Yang et al., 2020). El protocolo incluirá los siguientes pasos: (a) relleno hacia adelante con límite de cinco días hábiles consecutivos, que simula el valor de cierre vigente en días sin negociación; (b) eliminación de columnas con más del 30 % de valores faltantes sobre el total del período, garantizando continuidad histórica mínima; (c) interpolación lineal de los huecos residuales; y (d) cálculo de retornos logarítmicos diarios, que constituyen la variable de entrada al entorno de simulación. A través de esta técnica se busca consolidar la matriz de datos que haga posible identificar continuidad de las series, los niveles de liquidez y la presencia de datos faltantes, lo cual será correspondiente con los criterios definidos en los objetivos específicos. Metodológicamente este caso, Yang et al. (2020) indican que para desarrollar su trabajo incluyeron una etapa de reprocesamiento de datos antes de la evaluación de los agentes, mientras que Liu et al. (2020) esta etapa es crucial para lograr marcos y conjuntos de datos que sean comparables.

En tercer lugar, se empleará la simulación computacional con backtesting como técnica para obtener datos experimentales. Si bien los datos de entrada serán a partir de los históricos, la ejecución del entorno de gestión de portafolios se espera, generará nuevos registros asociados al comportamiento de los agentes PPO, A2C y DDPG y del ensamble propuesto. Es así como, esta técnica permitirá recolectar los resultados por ventana de entrenamiento, validación y prueba, así como registrar el desempeño de cada estrategia bajo los costos de transacción y las reglas de control de riesgo (Yang et al., 2020).

Finalmente, se utilizará una técnica de registro comparativo de resultados, orientada a documentar de forma ordenada las salidas del experimento para cada agente, el ensamble y las estrategias de referencia. A través de esta técnica se consolidarán los valores necesarios para el análisis posterior, tales como retorno acumulado, razón de Sharpe,

volatilidad anualizada, máximo drawdown y demás indicadores que se definan en el plan de análisis. Yang et al. (2020) lo hacen a partir de comparar el desempeño de su estrategia con los agentes individuales, el índice Dow Jones y la estrategia de mínima varianza mediante la razón de Sharpe. De este modo, las técnicas de recolección no solo permitirán obtener la información base del mercado, sino también reunir la evidencia empírica necesaria para contrastar el desempeño del modelo propuesto.

Instrumentos de recolección.

Recordando que, esta investigación es bajo un enfoque cuantitativo y con base en series históricas del mercado accionario, los instrumentos de recolección corresponderán a herramientas de técnicas de captura, organización y registro de información financiera y de resultados experimentales. De esta manera, se establece de forma preliminar los siguientes instrumentos de recolección.

- **Matriz de extracción.**

La matriz de extracción será diseñada para el registro por activo y por fecha de las variables como precio de cierre, precio ajustado, volumen negociado, rendimiento diario y demás series necesarias para la modelación. Cómo ya se mencionó, la comparación es fundamental, luego su paso previo debe garantizar trazabilidad, orden cronológico y comparabilidad entre activos. La matriz trabajará con frecuencia diaria y permitirá la identificación de datos faltantes, discontinuidades y diferencias de liquidez. Este instrumento se justifica porque tanto FinRL como la estrategia de ensamble de Yang et al. (2020) parten de conjuntos de datos de mercado previamente organizados y preprocesados antes del entrenamiento de los agentes.

- **Depuración de datos.**

El instrumento de depuración de datos funcionará a partir de los scripts que permitan la transformación y preparación los mismos. Estos serán desarrollados en un entorno de

programación para la estandarización de fechas, el cálculo de rendimientos, la depuración de valores faltantes y la construcción de las variables de entrada del modelo. Aunque se trata de código y no de un formulario convencional, en esta investigación funciona como instrumento porque permite, mediante el ejercicio de reprocesamiento, convertir la información histórica en una base utilizable para el entrenamiento y la evaluación.

- **Simulador de Gestión de Portafolios.**

En el entorno computacional entra el simulador de gestión de portafolio, el cual permitirá definir el estado del mismo entorno, la acción del agente, los costos de transacción, la acción del agente y la regla de control de riesgo por turbulencia o régimen. Este instrumento es el centro operativo del presente estudio, ya que permitirá recolectar los resultados producidos por la interacción de los agentes PPO, A2C y DDPG en el mercado simulado. La pertinencia de este instrumento está respaldada por FinRL, que trabaja con entornos virtuales configurados con datos del mercado y evaluados mediante backtesting, y por estudios como el de Jiang, Olmo y Atwi (2024), que incorporan restricciones de riesgo y costos de transacción en la función de recompensa del entorno.

Así las cosas, se implementará un entorno personalizado de gestión de portafolios siguiendo la especificación de la biblioteca Gymnasium (gym.Env), que actúa como el instrumento central del experimento. Este instrumento recibe los datos históricos procesados y simula la operación del portafolio bajo las condiciones que se describen a continuación.

  - **Espacio de acción y función de recompensa.**

El espacio de acción es continuo y está definido mediante softmax sobre un vector de señales reales, lo que garantiza que los pesos sumen uno y sean no negativos en todo momento. La función de recompensa se define como el logaritmo natural del retorno neto del portafolio después de costos de transacción:

$$r_t = \log(1 + retornoportafolio - costostransacción)$$

El uso del logaritmo penaliza asimétricamente las pérdidas grandes y desincentiva la toma de riesgos excesivos, en línea con los marcos teóricos de utilidad logarítmica (Sharpe, 1994).

- **Costos de transacción.**

Los costos se modelan como proporcionales al cambio en los pesos del portafolio, calibrados al 0.1 % por operación, nivel representativo del mercado bursátil colombiano considerando spreads bid-ask y comisiones de las firmas comisionistas (Deng et al., 2017). Esta parametrización evita que el modelo aprenda estrategias con rotación excesiva e irrealista.

- **Mecanismo de control de riesgo por turbulencia.**

Cuando el índice de turbulencia financiera de Kritzman y Li (2010), calculado como la distancia de Mahalanobis entre el retorno del día actual y la distribución histórica de retornos de los últimos 252 días hábiles, supera el percentil 90 de la ventana de entrenamiento, el portafolio asume composición cero (posición defensiva). El umbral se calcula exclusivamente con datos de la ventana de entrenamiento para evitar look-ahead bias.

- **Plantilla de Registro de Resultados.**

Este instrumento está orientado a documentar el desempeño de cada agente, del ensamble y de las estrategias de referencia en cada ventana de validación y prueba. Esta plantilla incluirá métricas como el retorno acumulado, la razón de Sharpe, volatilidad anualizada, máximo drawdown y rotación del portafolio. Así, se tendrá el registro comparativo de resultados de cada ventana de prueba, los cuales estarán en una estructura de datos que consolida para cada agente y para el ensamble las siguientes métricas: retorno acumulado, razón de Sharpe —calculada con tasa libre de riesgo del 6 % anual correspondiente a los Títulos de Tesorería (TES) colombianos—, volatilidad anualizada,

máximo drawdown y ratio de Calmar. Esta información se exporta al final del experimento en formato Excel con cinco hojas diferenciadas, siguiendo el esquema comparativo utilizado por Yang et al. (2020).

Procedimiento.

El procedimiento de la investigación se desarrollará en varias etapas que serán secuenciales. En una primera instancia, se definirá el conjunto de activos que harán parte del estudio con base en criterios de continuidad histórica, liquidez relativa y disponibilidad de información. A partir de esta delimitación, se realizará la extracción y consolidación de los datos históricos del mercado accionario colombiano en el instrumento matriz diseñado para tal fin. En esta fase, es donde se incluirán las variables ya mencionadas como precios de cierre, precios ajustados, volúmenes negociados y rendimientos, entre otras. Posteriormente, se ejecutará un proceso de depuración y preparación de la base de datos, enfocado también en identificar valores faltantes, homogeneizar fechas, estandarizar formatos y construir las variables necesarias para el entrenamiento de los modelos.

En una segunda etapa se configurará el entorno computacional de simulación para la gestión de portafolios. Aquí se definirán los estados del sistema, las acciones de asignación, la función de recompensa, los costos de transacción y la regla de control de riesgo por turbulencia o régimen. Una vez configurado el entorno, se procederá al entrenamiento de los agentes Proximal Policy Optimization (PPO), Advantage Actor-Critic (A2C) y Deep Deterministic Policy Gradient (DDPG), bajo condiciones homogéneas de simulación y con una misma estructura de datos. Después, se aplicará el esquema de ventanas móviles de entrenamiento, validación y prueba, de manera que en cada iteración los agentes sean evaluados en una ventana de validación y se seleccione para la siguiente ventana de prueba aquel que presente el mejor desempeño ajustado por riesgo medido mediante la razón de Sharpe.

En una tercera etapa se registrarán y consolidarán los resultados experimentales generados por cada agente, teniendo en cuenta el ensamble y las estrategias de referencia. Esta información se organizará en la plantilla de registro de resultados que permitirá documentar de forma comparable indicadores como el retorno acumulado, la volatilidad anualizada, razón de Sharpe, máximo drawdown y rotación del portafolio. Como controles técnicos, se evitará el uso de información futura dentro del entrenamiento, se mantendrán las reglas uniformes para todos los agentes y se documentarán los parámetros utilizados en cada corrida del experimento.

Validación de instrumentos.

La validación de los instrumentos se realizará en dos niveles: revisión de pertinencia técnica y metodológica, y prueba piloto con datos reales del mercado colombiano.

- **Revisión de Pertinencia Técnica y Metodológica**

Cada instrumento fue evaluado respecto a su coherencia con los objetivos de la investigación y con los estándares metodológicos de la literatura de DRL aplicado a finanzas (Yang et al., 2020; Jiang et al., 2024). La validación cubrió tres instrumentos principales. La matriz de extracción fue validada cruzando precios de cierre con fuentes alternativas para fechas de alta bursatilidad, y verificando que el tratamiento del formato numérico de Investing.com —punto como separador de miles y coma como separador decimal— produjera valores correctos. Los scripts de depuración fueron verificados confirmando que el relleno hacia adelante no superara en ningún caso los cinco días hábiles, y que la interpolación no introdujera valores fuera del rango histórico de cada activo.

El entorno de simulación fue validado técnicamente en tres aspectos: (a) verificación de que los pesos del portafolio suman 1.0 en cada paso de tiempo con precisión numérica; (b) confirmación de que los costos de transacción se aplican correctamente al cambio en la composición del portafolio; y (c) comprobación de que el índice de turbulencia se calcula

exclusivamente con datos de la ventana de entrenamiento, sin filtración de información futura al proceso de decisión.

- **Prueba Piloto**

Se ejecutó una corrida completa del experimento con TIMESTEPS = 20.000 pasos de entrenamiento sobre el período 2015–2025. Esta prueba permitió identificar y corregir los siguientes aspectos antes de la corrida definitiva.

En cuanto a la funcionalidad del sistema, se confirmó que el sistema de checkpoints permite retomar una corrida interrumpida desde la última ventana completada, y que el EvalCallback selecciona y guarda el modelo con mejor desempeño en validación en lugar del modelo del último paso de entrenamiento. En cuanto a las fuentes de datos, se identificó que cuatro tickers —PFBCOLOM, ECOPETROL, GRUPOARGOS y GRUPOSURA— requieren carga desde CSV de Investing.com debido a cobertura insuficiente en las fuentes automáticas. En cuanto a las características del universo, se confirmó que la correlación promedio entre los 14 activos seleccionados es de 0.2, nivel que garantiza beneficios de diversificación relevantes para la gestión del portafolio (Markowitz, 1952, citado en Yang et al., 2020).

Plan de análisis.

El análisis de la información se desarrollará en varias fases articuladas entre sí. En una primera fase se realizará el análisis descriptivo de la base de datos construida, con el fin de caracterizar los activos seleccionados en términos de continuidad histórica, liquidez, comportamiento de precios y presencia de datos faltantes. Aquí se verificará la consistencia de la información utilizada y establecer las condiciones iniciales del entorno de simulación. Además, servirá para justificar la selección final de los activos que harán parte del estudio y para documentar las decisiones de depuración y preparación de los datos.



En la fase dos se llevará a cabo el análisis experimental del desempeño de los modelos. Para ello, se entrenarán los agentes PPO, A2C y DDPG dentro del entorno de gestión de portafolios y se ejecutará el ensamble mediante el esquema de ventanas móviles definido en el diseño metodológico. Como ya se ha mencionado en las secciones anteriores, los resultados obtenidos en cada ventana de validación y prueba serán registrados para identificar el comportamiento de cada agente y del ensamble bajo condiciones cambiantes del mercado. A partir de esta ejecución se calcularán indicadores de desempeño como retorno acumulado, razón de Sharpe, volatilidad anualizada, máximo drawdown y rotación del portafolio, los cuales permitirán evaluar la relación entre rentabilidad, riesgo y estabilidad del modelo propuesto.

En una tercera fase se realizará un análisis comparativo entre el ensamble, los agentes individuales y las estrategias de referencia. El propósito será establecer si el ensamble presenta ventajas en términos de desempeño ajustado por riesgo frente a las alternativas evaluadas. Para ello, se examinarán las diferencias observadas en las métricas definidas, como retorno acumulado, razón de Sharpe, volatilidad anualizada y máximo drawdown, y

se identificarán patrones de comportamiento por ventana o por régimen de mercado. Este análisis permitirá valorar no solo qué estrategia obtiene mejores resultados, sino también cuál muestra un comportamiento más estable y consistente bajo condiciones cambiantes del mercado.

En la fase cuatro, se desarrollará un análisis de sensibilidad y robustez, orientado a examinar cómo cambian los resultados cuando se modifican parámetros clave del experimento. Entre estos el tamaño de las ventanas de entrenamiento, validación y prueba, el nivel de costos de transacción y las reglas de control de riesgo por turbulencia o régimen. Esta fase permitirá determinar si el desempeño del ensamble mantiene un comportamiento relativamente estable o no bajo escenarios alternativos.

Resultados preliminares de investigación

Los resultados que se presentan a continuación corresponden a la corrida definitiva del experimento (TIMESTEPS = 100.000) y cubren las cuatro fases del plan de análisis: análisis descriptivo de la base de datos, desempeño experimental de los modelos, análisis comparativo frente a benchmarks y análisis de sensibilidad a costos de transacción.

- **Análisis Descriptivo de la Base de Datos.**

El universo de inversión quedó conformado por 14 activos del índice COLCAP con cobertura histórica suficiente para el período enero 2015–diciembre 2025. Los precios normalizados en base 100 evidencian marcada heterogeneidad en el desempeño acumulado: activos como PFDAVVNDA y CNEC registran crecimientos superiores al 1.000 % en el período, mientras que la mayoría de los activos se mantiene en un rango de 100–300. Se observa un quiebre estructural pronunciado en el primer trimestre de 2020, correspondiente al shock derivado de la pandemia de COVID-19, que deprimió

simultáneamente todos los activos del universo. A partir de 2023–2024 se registra una fase de recuperación diferenciada según sector.

La matriz de correlaciones de retornos diarios revela un patrón de correlación moderada. Los coeficientes más altos se observan entre GRUPOARGOS y GRUPOSURA (r = 0.5), y entre PFBCOLOM y ambas empresas del Grupo (r = 0.4), lo que refleja vínculos corporativos en el conglomerado empresarial colombiano. La mayoría de los pares presenta correlaciones inferiores a 0.3, confirmando que el universo ofrece beneficios de diversificación relevantes para la gestión del portafolio.

- **Índice de Turbulencia Financiera**

El índice de turbulencia financiera calculado según Kritzman y Li (2010) presenta un umbral de referencia global de p90 = 31.7 sobre el período completo. Los episodios de estrés identificados son consistentes con eventos documentados del mercado colombiano y global: el pico más alto (aproximadamente 2.100 unidades de distancia de Mahalanobis) se registra en 2018; el clúster de 2020 corresponde al crash asociado a la pandemia; los picos de 2022 reflejan el ciclo alcista de tasas de interés globales; y los eventos de 2024–2025 coinciden con la incertidumbre política y fiscal en Colombia. La baja frecuencia de superación del umbral p90 confirma que el mecanismo de control de riesgo se activa selectivamente, sin penalizar en exceso los períodos de mercado normal.

- **Desempeño del Ensamble Frente a Benchmarks**

La evolución del portafolio en base 100 para el período de prueba fuera de muestra (2018–2026) muestra los siguientes resultados. El ensamble DRL alcanzó un valor aproximado de 480 (retorno acumulado del 380 %) al final del período, superando ampliamente al COLCAP real (retorno aproximado del 20 %) y a la estrategia de igual ponderación —equally weighted, EW— (retorno aproximado del 90 %). El algoritmo DDPG fue el de mayor desempeño individual, con un valor cercano a 780 al cierre del período,

aunque con mayor volatilidad intraperiodo, especialmente en el repunte de 2024–2025. PPO y A2C mostraron desempeños más moderados, similares entre sí, con valores entre 120 y 160 al cierre del período piloto. El ensamble captura aproximadamente el 60 % del retorno de DDPG con una trayectoria más estable, lo que evidencia su función de amortiguación del riesgo en el marco propuesto por Yang et al. (2020).

- **Análisis del Drawdown Comparativo**

La comparación de los drawdowns del ensamble DRL, COLCAP y EW en el período completo muestra que el crash del primer trimestre de 2020 generó caídas de hasta el 45 % en el COLCAP real y de hasta el 35 % en la estrategia EW. El ensamble DRL limitó la caída a aproximadamente el 12 %, resultado que constituye la ventaja más visible del mecanismo de control de riesgo por turbulencia implementado con base en Kritzman y Li (2010). En los períodos 2022–2023 y 2023–2024, el ensamble mostró drawdowns máximos en el rango del 8 % al 22 %, frente al 28 %–35 % del COLCAP. La recuperación del ensamble desde los valles es consistentemente más rápida que la de los benchmarks pasivos.

- **Dinámica de Selección de Agentes**

El agente DDPG fue seleccionado como activo en 23 de las 29 ventanas de validación (79 %), seguido por A2C en cuatro ventanas (14 %) y PPO en dos ventanas (7 %). Temporalmente, DDPG domina de manera casi continua desde 2018 hasta 2026, con breves interrupciones de A2C en 2019–2020 y en 2023. PPO fue seleccionado únicamente en una ventana de 2021 y en la última de 2025. Este patrón sugiere que la naturaleza determinística del DDPG, combinada con su memoria de largo plazo —replay buffer de 50.000 pasos— (Lillicrap et al., 2016), le confiere ventaja en el horizonte de validación de tres meses del mercado colombiano. Sin embargo, este resultado podría ser parcialmente

sensible al número de pasos de entrenamiento de la versión piloto, por lo que se reserva la interpretación definitiva para la corrida con 100.000 pasos.

- **Razón de Sharpe por Ventana de Prueba**

La evolución de la razón de Sharpe para cada agente en cada ventana de prueba evidencia alta variabilidad entre ventanas para todos los agentes, con un rango entre -5 y +8.5, lo que confirma la no estacionariedad del mercado colombiano y la pertinencia del esquema de selección dinámica adoptado. DDPG registra los valores extremos más altos (pico máximo de aproximadamente 8.5 en 2024) y también algunos valles profundos en 2022–2023. PPO tiende a ser el agente con mayor estabilidad relativa, aunque raramente el de mejor desempeño. Las ventanas con razón de Sharpe negativa para todos los agentes —particularmente en 2020 y 2022— coinciden con los episodios de estrés identificados en el índice de turbulencia, validando la coherencia interna del marco metodológico.

- **Análisis de Sensibilidad a Costos de Transacción**

La tabla de sensibilidad evalúa el impacto de cinco niveles de costo de transacción sobre las métricas principales del ensamble. Cabe señalar que este análisis constituye una aproximación basada en un turnover promedio del 5 %, ya que el turnover calculado directamente desde el entorno de simulación no está disponible en la versión piloto.

Costo tx (%)	Sharpe	Retorno anual (%)	Máx. drawdown (%)
0.01	2.893	100.2	−12.3
0.05	2.885	99.8	−12.4
0.10	**2.874**	**99.3**	**−12.5**
0.20	2.851	98.3	−12.6
0.50	2.783	95.3	−12.9

Nota. Resultados de la corrida piloto (TIMESTEPS = 20.000). Los valores de Sharpe y retorno anual se calcularon con tasa libre de riesgo del 6 % anual (TES Colombia). El

análisis de sensibilidad asume un turnover promedio constante del 5 %, lo que constituye una aproximación que se corregirá en la corrida definitiva.

La razón de Sharpe disminuye de manera aproximadamente lineal desde 1.58 con costos mínimos hasta 1.23 en el escenario más conservador (0.50 %), lo que indica que el ensamble no opera con rotación excesiva. El retorno anual cae aproximadamente 9.5 puntos porcentuales bajo el escenario de mayor costo, reducción material pero que no elimina la ventaja sobre los benchmarks. El máximo drawdown empeora moderadamente, de −29 % a −33.5 %, confirmando que el riesgo a la baja no es muy sensible al nivel de costos de negociación. Dado que los costos efectivos del mercado colombiano oscilan típicamente entre 0.05 % y 0.20 % por operación, la estrategia mantiene razones de Sharpe superiores a 1.4 en el rango de costos plausible.

Métricas comparativas de desempeño – Corrida definitiva (TIMESTEPS = 100.000)

Estrategia	Ret. acum.	Sharpe	Vol. anual	Máx. DD	Calmar	Período
Ensamble DRL	2,595%	2.897	22.9%	−12.3%	8.17	2018–2026
DDPG	2,934%	2.803	24.7%	−12.3%	8.58	2018–2026
PPO	748%	2.769	14.5%	−17.2%	3.32	2018–2026
A2C	20.6%	−0.041	16.6%	−39.6%	0.10	2018–2026
COLCAP (real)	21.2%	−0.053	20.6%	−46.7%	0.06	2018–2026
Igual pond. (EW)	90.3%	0.242	17.9%	−42.0%	0.22	2018–2026

Nota. DD = drawdown. rf = tasa libre de riesgo (6 % anual, TES Colombia). EW = estrategia de igual ponderación con rebalanceo diario. COLCAP (real) = índice descargado desde Investing.com en formato CSV. Sharpe calculado como exceso de retorno diario sobre rf/252, anualizado por √252. Calmar = retorno anualizado / |máximo drawdown|. Resultados de la corrida definitiva (TIMESTEPS = 100.000).

Conclusiones Preliminares

- **Viabilidad del Enfoque**

La corrida definitiva del experimento confirma que es técnicamente viable implementar un ensamble de agentes DRL —PPO, A2C y DDPG— para la gestión de portafolios en el mercado accionario colombiano utilizando datos reales del período 2015–2025, con 100.000 pasos de entrenamiento por agente por ventana. El entorno de simulación operó correctamente bajo los parámetros especificados, el mecanismo de selección dinámica funcionó sin look-ahead bias y el sistema de checkpoints garantizó la reproducibilidad del experimento ante interrupciones de la sesión de cómputo.

- **Desempeño del Ensamble**

Los resultados definitivos muestran que el ensamble DRL supera a todos los benchmarks pasivos tanto en retorno acumulado como en desempeño ajustado por riesgo. El ensamble alcanzó un retorno acumulado del 2.595 % y una razón de Sharpe de 2.897 (rf = 6 % anual) en el período de prueba fuera de muestra julio 2018 – septiembre 2026, frente al 21.2 % y Sharpe de −0.053 del COLCAP real, y el 90.3 % y Sharpe de 0.242 de la estrategia de igual ponderación. La ventaja es estadísticamente significativa al 1 % frente al COLCAP ($t = +5.007$, $p < 0.001$) y frente a la estrategia EW ($t = +5.290$, $p < 0.001$). El máximo drawdown del ensamble fue de −12.3 %, frente al −46.7 % del COLCAP real y el −42.0 % de EW, lo que confirma la efectividad del mecanismo de control de riesgo por turbulencia.

- **Dinámica de Selección de Agentes**

La dominancia de DDPG como agente activo en el 79 % de las ventanas (23 de 29) es coherente con las características del mercado colombiano: un mercado con tendencias moderadas y alta concentración en pocos sectores, contexto en el que la memoria de largo plazo del replay buffer de DDPG ofrece una ventaja sistemática frente a métodos sin memoria como A2C (Lillicrap et al., 2016). Sin embargo, la activación de A2C en períodos

de alta turbulencia —pandemia 2020 y crisis bancaria 2023— confirma que ningún agente es óptimo en todos los regímenes de mercado, validando la hipótesis central del ensamble adaptativo. El ensamble no supera a DDPG de forma estadísticamente significativa ($p = 0.5753$) porque ambas curvas comparten el 79 % de la trayectoria, resultado que era esperable dado el diseño del mecanismo de selección.

- **Robustez ante Costos de Transacción**

El análisis de sensibilidad confirma que la estrategia mantiene razones de Sharpe superiores a 2.78 incluso bajo el escenario de mayor costo modelado (0.50 %). En el rango de costos plausible para el mercado colombiano (0.05 %–0.20 %), el ensamble opera con Sharpe entre 2.85 y 2.89, representando una ventaja robusta y sostenida sobre los benchmarks. La caída de solo 0.11 puntos de Sharpe entre el escenario de menor y mayor costo confirma que el modelo no opera con rotación excesiva, resultado consistente con la penalización por costo de transacción incorporada en la función de recompensa.

- **Limitaciones de los Resultados Preliminares**

Los resultados deben interpretarse con las siguientes reservas. Primero, la selección del universo de activos con la composición actual del COLCAP introduce sesgo de supervivencia, dado que los activos deslistados antes de 2025 no están representados en la muestra. Segundo, las pruebas estadísticas basadas en t-test de muestras independientes asumen independencia entre observaciones, supuesto que no se cumple estrictamente para series temporales con estructura autocorrelacionada. Tercero, el análisis de sensibilidad al tamaño de las ventanas de entrenamiento y la estrategia de mínima varianza como benchmark adicional no se ejecutaron en la versión definitiva. Cuarto, la liquidez real de algunos activos del COLCAP puede diferir de los supuestos de la simulación.

Recomendaciones preliminares

Hay tres pasos para completar el experimento: en primer lugar, aumentar TIMESTEPS de 20.000 a 100.000 pasos de entrenamiento, cambio ya previsto en el código como la configuración de producción. En segundo lugar, implementar la estrategia de mínima varianza como benchmark adicional, calculada mediante optimización cuadrática sobre los retornos históricos de cada ventana de entrenamiento, en línea con el esquema comparativo de Yang et al. (2020). En tercer lugar, agregar el cálculo explícito del turnover por ventana dentro del entorno de simulación, acumulando el valor absoluto del cambio de pesos en cada paso de tiempo y exponiéndolo como métrica de salida. En cuarto lugar, implementar el análisis de sensibilidad al tamaño de ventanas ejecutando el loop con al menos tres configuraciones -por ejemplo, 24/3/3, 36/3/3 y 48/3/3 meses- y comparando las métricas resultantes.

Hasta el momento, los resultados preliminares sugieren al menos tres líneas de extensión. La primera consiste en explorar mecanismos de ensamble alternativos a la selección por mejor Sharpe en validación, como una votación ponderada por desempeño histórico o una mezcla de distribuciones de acción, que podrían capturar la complementariedad entre agentes de manera más continua. La segunda es la de incorporar variables macroeconómicas colombianas como tasa de cambio, precio del petróleo Brent, tasa de política monetaria del Banco de la República, entre otras, como variables de estado adicionales del entorno. Finalmente, se puede extender el análisis a otros mercados de la región, por ejemplo, al IPC México, o al Bovespa Brasil para evaluar si la dominancia de DDPG es específica del mercado colombiano o es generalizable a mercados emergentes con características similares.

Contribuciones originales esperadas

La principal contribución esperada de esta investigación consiste en adaptar y evaluar un esquema de ensamble con aprendizaje por refuerzo profundo para la gestión de portafolios en el mercado accionario colombiano, un contexto que ha sido poco explorado en la literatura frente a mercados más líquidos y desarrollados. En este sentido, el proyecto busca aportar evidencia empírica sobre el comportamiento de un ensamble basado en PPO, A2C y DDPG en un mercado emergente caracterizado por menor liquidez relativa, mayores fricciones operativas y mayor sensibilidad a cambios de régimen. Esta adaptación constituye un aporte metodológico, porque no se limita a replicar modelos existentes, sino que los ajusta a condiciones reales del entorno bursátil colombiano mediante ventanas móviles, costos de transacción y control de riesgo por turbulencia.

De manera complementaria, la investigación aspira a generar una contribución aplicada para el campo de las finanzas cuantitativas y la analítica financiera en Colombia, al proponer un protocolo reproducible para la construcción de datos, entrenamiento de agentes, validación y comparación de estrategias. A partir de los avances preliminares del experimento, se espera que el estudio sirva como punto de partida para futuras aplicaciones de inteligencia artificial en la gestión de portafolios y en la evaluación de estrategias de inversión bajo criterios de riesgo y robustez. Así, el aporte original del proyecto no solo se ubica en el plano académico, sino también en su potencial utilidad para fortalecer capacidades analíticas en el contexto financiero colombiano y en otros mercados emergentes con características similares.

Cronograma

A continuación, en la Tabla 2, detalle las actividades a realizar para alcanzar cada uno de los objetivos propuestos. Debe indicar en tiempo (meses) cuánto tardaría en desarrollar cada actividad (entre 5 y 9 meses).

Tabla 2. Plantilla para el planteamiento del cronograma de trabajo.

| | **Actividades a desarrollar** | **Semanas** | **Tiempo del proyecto en meses (5-9 meses)** | | |

| | | |

Referencias

Bolsa de Valores de Colombia - BVC. (s.f.). *Mercado de Valores.* Obtenido de https://www.bvc.com.co/mercado-de-valores?tab=leyes

Congreso de la República de Colombia. (2005). *Ley 964 de 2005. Por la cual se dictan normas generales y se señalan en ellas los objetivos y criterios a los cuales debe sujetarse el Gobierno Nacional para regular las actividades de manejo, aprovechamiento e inversión de recursos captados del público.* Obtenido de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=22412

Deng, Y., Feng, B., Kong, Y., Ren, Z., & Dai, Q. (2017). Deep Direct Reinforcement Learning for Financial Signal Representation and Trading. *IEEE Transactions on Neural Networks and Learning Systems, 28*(3), 653-664. Obtenido de https://par.cse.nsysu.edu.tw/resource/paper/2017/170801/Deep%20Direct%20Reinforcement%20Learning%20for%20Financial%20Signal%20Representation%20and%20Trading%20.pdf

Jiang, Y., Olmo, J., & Atwi, M. (2024). Deep reinforcement learning for portfolio selection. *Global Finance Journal, 62*. Obtenido de https://www.sciencedirect.com/science/article/pii/S1044028324000887

Jiang, Z., Xu, D., & Liang, J. (2017). *A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem.* Obtenido de https://arxiv.org/abs/1706.10059

Kritzman, M., & Li, Y. (2010). Skulls, Financial Turbulence and Risk Management. *Financial Analyst Journal, 66*(5). Obtenido de https://www.tandfonline.com/doi/abs/10.2469/faj.v66.n5.3

Lee, J., Kim, R., Yi, S.-W., & Kang, J. (2020). *MAPS: Multi-agent Reinforcement Learning-based Portfolio Management System.* Obtenido de https://www.ijcai.org/proceedings/2020/0623.pdf

Lillicrap, T., Hunt, J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., . . . Wierstra, D. (2016). *Continuous control with Deep Reinforcement Learning.* Obtenido de https://arxiv.org/pdf/1509.02971

Liu, X.-Y., Yang, H., Chen, Q., Zhang, R., Yang, L., Xiao, B., & Wang, C. D. (2020). *FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance.* Obtenido de https://arxiv.org/pdf/2011.09607

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance, 7*(1), 77-91. Obtenido de https://www.jstor.org/stable/2975974

Mejía Estrada, D. (2023). *Aprendizaje reforzado profundo para la administración de portafolios de renta fija (Tesis de Maestría).* Universidad EAFIT. Obtenido de https://repository.eafit.edu.co/entities/publication/5ebe26a7-431a-46b0-a4dc-c52059931f92

Mnih, V., Puigdoménech Badia, A., Mirza, M., Graves, A., Harley, T., Lillicrap, T., . . . Kavukcuoglu, K. (2016). *Asynchronous Methods for Deep Reinforcement Learning.* Conference on Machine Learning. Obtenido de https://arxiv.org/pdf/1602.01783

Obando Morales, S. (2022). *Modelo de aprendizaje profundo reforzado aplicado al trading de Bitcoin (Tesis de maestría).* Universidad EAFIT. Obtenido de https://repository.eafit.edu.co/entities/publication/ffa13d31-433a-4189-baf6-41637d377e0c

Presidencia de la República de Colombia. (2010). *Decreto 2555 de 2010. Por el cual se recogen y reexpiden las normas en materia del sector financiero, asegurador y del mercado de valores y se dictan otras disposiciones.* Obtenido de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=40032

Presidencia de la República de Colombia. (2016). *Decreto 1247 de 2016. Por el cual se sustituye el Título 7 del Libro 9 de la Parte 2 del Decreto 2555 de 2010 la administración de portafolios de terceros.* Obtenido de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php

Presidencia de la República de Colombia. (2018). *Decreto 661 DE 2018. Por el cual se modifica el Decreto 2555 de 2010 en lo relacionado con la actividad de asesoría y se dictan otras disposiciones.* Obtenido de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=85921

Presidencia de la República de Colombia. (2024a). *Decreto 0265 de 2024.* Obtenido de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=234890

Presidencia de la República de Colombia. (2024b). *Decreto 1239 de 2024. Por el cual se modifica el Decreto 2555 de 2010 en lo relacionado con los diferentes instrumentos que contribuyen a la liquidez del mercado de valores y se dictan otras disposiciones.* Obtenido de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=253417

Quintero González, C. (2024). *Aprendizaje reforzado en pair-trading: Aplicación para una estrategia pair-trading.* ODEON. Obtenido de https://revistas.uexternado.edu.co/index.php/odeon/article/view/10070

Schulman, J., Wolski, F. D., Radford, A., & Klimov, O. (2017). *Proximal Policy Optimization Algorithms.* Obtenido de https://arxiv.org/pdf/1707.06347

Sharpe, W. (1994). The Journal of Portfolio Management. *The Journal Portfolio Management, 21*(1). Obtenido de https://www.pm-research.com/content/iijpormgmt/21/1/49

Sutton, R., & Barto, A. (2018). *Reinforcement Learning. An Introduction.* MIT Press. Obtenido de https://mitpress.mit.edu/9780262039246/reinforcement-learning/

Yang, H., Zhong, S., Liu, X.-Y., & Walid, A. (2020). *Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy.* Obtenido de https://openfin.engineering.columbia.edu/sites/openfin.engineering.columbia.edu/files/content/publications/ensemble.pdf

Sugerencia (1) de director de trabajo de grado:	
Información del docente	Nombre: Jaime Alberto Ochoa Durán
	Correo institucional: jaochoa@universidadean.edu.co
Sugerencia (2) de director de trabajo de grado:	
Información del docente	Nombre: William Zuluaga Muñoz
	Correo institucional: WZULUAGA@universidadean.edu.co

Tenga en cuenta que esto es una sugerencia para el Área de Trabajos de Grado. Se debe verificar la disponibilidad del docente.