Aprendizaje Reforzado Profundo para la Administración de Portafolios de Renta Fija

Deep Reinforcement Learning for Automated Fixed Income Portfolio Management

Estudiante

David Mejía Estrada

dmejiae3@eafit.edu.co

Asesora

Paula María Almonacid Hurtado

palmona1@eafit.edu.co

UNIVERSIDAD EAFIT

Escuela de Administración

Maestría en Ciencias de los Datos y Analítica

Medellín

2023

Contenido

Listado de Figuras

Listado de tablas

Resumen

En este trabajo se aplican técnicas de aprendizaje reforzado profundo en la administración de portafolios de inversión de renta fija, específicamente títulos soberanos emitidos por el gobierno colombiano. El periodo de análisis comprende siete años, desde enero de 2015 hasta diciembre de 2022. Encontramos que es posible generar rentabilidad y lograr una eficiente gestión del riesgo como resultado de las estrategias de "trading" que los modelos de aprendizaje reforzado profundo prevén más convenientes dadas ciertas condiciones de mercado y de cada uno de los títulos, como su riesgo implícito en métricas como DV01, Duración y Convexidad. Finalmente, este estudio contribuye al campo de las aplicaciones de aprendizaje de máquina e inteligencia artificial sobre administración de carteras de inversión, con un enfoque relativamente nuevo sobre el mercado de renta fija en general, consolidándose como uno de los primeros trabajos en aplicar técnicas de aprendizaje por refuerzo al mercado de deuda pública colombiana.

Abstract

This paper applies deep reinforced learning techniques to the management of fixed income investment portfolios, specifically sovereign securities issued by the Colombian government. The period of analysis covers seven years, from January 2015 to December 2022. We find that it is possible to generate profitability and achieve efficient risk management because of the trading strategies that deep reinforced learning models foresee more convenient given certain market conditions and of each of the securities, such as their implied risk in metrics like DV01, Duration and Convexity. Finally, this study contributes to the field of machine learning and artificial intelligence applications on investment portfolio management, with a relatively new focus on the fixed income market in general, consolidating itself as one of the first works to apply reinforcement learning techniques to the Colombian public debt market.

Palabras Clave: Yield curve; Machine Learning; Trading strategy; Deep Reinforcement Learning; Fixed Income; Risk Management; Portfolio Management.

1. INTRODUCCIÓN

1.1. Planteamiento del problema

Las estrategias de trading sobre portafolios de renta fija desempeñan un papel fundamental en la gestión de activos y pasivos de las entidades financieras y de cualquiera que gestione este tipo de portafolios de manera independiente[1]. En todo momento, estos agentes intentan rentabilizar sus portafolios al tiempo que intentan minimizar el riesgo de tasa de interés asumido, utilizando para ello la información que puede entregar la curva de rendimientos y sus hipotéticos movimientos futuros -como su aplanamiento o empinamiento- así como la posible variación de las tasas de interés y precios de mercado de los títulos que componen la curva vistos de manera individual. Todos estos agentes se encuentran en los mercados financieros de deuda, con objetivos diferentes respecto a la gestión de sus activos y pasivos, siendo los mercados de deuda soberana -este es, donde se negocian los títulos de deuda emitidos por el gobierno de un país para financiar sus necesidades de gasto público- uno de los elegidos por estos agentes para esa labor. En este trabajo, se plantea trabajar con el mercado de deuda soberana de Colombia, lugar donde confluyen diversos agentes con objetivos diferentes.

El mercado de bonos soberanos es un mercado financiero importante, amplio y líquido en Colombia, donde bancos, instituciones e inversionistas extranjeros y privados negocian todos los días grandes cantidades de esos bonos emitidos por el gobierno colombiano[2]. Usualmente, esas instituciones no sólo invierten en un título o tenor específico, sino que también tienen una gama de posibilidades para negociar en ese mercado, como, por ejemplo, apalancar ventas en corto con operaciones repo, y operaciones de corretaje, entre otras, que juegan un papel crítico en la gestión de activos y pasivos para esas empresas. Para ello, las instituciones financieras y los operadores independientes constituyen portafolios de inversión que tienden a valorizarse con el tiempo debido a los intereses que nocionalmente devengan estos títulos.

[1] El concepto de Asset and Liability Management (ALM) para las entidades financieras suele ser bastante relevante. Es a partir de este concepto que el negocio bancario puede ser rentable y sostenible en el tiempo, y cuenta con un alto nivel de complejidad técnico, ampliamente explicado en (Zenios & Ziemba, 2007).

[2] De acuerdo con cifras oficiales publicadas por la Bolsa de Valores de Colombia en (Bolsa de Valores de Colombia S.A, 2023), el monto diario negociado de deuda pública en el mercado público promedia los 1,8 billones de pesos colombianos, convirtiéndolo así en el mercado de valores más líquido, amplio y profundo que opera en Colombia. Si se desea, se puede comparar con el volumen medio del mercado accionario de tan solo 70 mil millones de pesos colombianos al día.

Sin embargo, los gestores de carteras pueden aumentar su rentabilidad negociando los bonos antes de su fecha de vencimiento, lo que implica que los portafolios de Renta Fija tendrán una exposición significativa al riesgo de mercado, en este caso, al riesgo de tasas de interés.

Los gestores de portafolios no sólo toman posiciones en un nodo concreto de la curva de rendimientos, sino que también lo hacen a lo largo de toda la curva. La estrategia depende de las expectativas sobre los movimientos de la curva. Por ejemplo, si el gestor espera que la curva de rendimientos se aplane -lo que significa que las tasas de corto plazo subirán mientras que las tasas de largo plazo caerán-, entonces tomará posiciones cortas en bonos a corto plazo y posiciones largas en bonos a largo plazo, y dicha estrategia tendrá un riesgo de mercado asociado.

Esta es una tarea compleja, porque tanto el movimiento de los precios y tasas de interés de los bonos vistos de manera individual como los movimientos de la curva de rendimientos completa son completamente erráticos, volátiles y no lineales, lo que supone una dificultad importante para hacer predicciones (Henrique y otros, 2019), y, por tanto, para tomar posiciones de trading a lo largo de la curva mientras también se busca minimizar el riesgo de mercado asumido.

1.2. Justificación

Zenios y Zeimba (2007), explican las carteras o portafolios de renta fija desempeñan un papel importante en la gestión de activos y pasivos, tanto para las instituciones financieras como para las no financieras, e incluso para otros fines y contextos. Sin embargo, el mercado de renta fija presenta cierta complejidad, especialmente en la gestión del riesgo de mercado. Los administradores de portafolios suelen tener que jugar con muchas variables dentro de su actuación, como la inflación, la política monetaria y la liquidez del mercado, entre otras. Esas variables cambian a lo largo del tiempo en patrones no lineales, y afectan a la formación de los precios de los títulos de renta fija, y, por tanto, de la curva de rendimientos.

Entonces, es demasiado importante para la industria financiera encontrar nuevas alternativas eficientes para la gestión de carteras de renta fija, utilizando tanto estrategias de curva como posiciones direccionales individuales. Los gestores de carteras buscan aumentar su

rentabilidad a la vez que reducen su riesgo de mercado, por lo que el objetivo principal es maximizar la relación entre la rentabilidad y el riesgo asumido.

En la actualidad, el Machine Learning ha ayudado a resolver algunos problemas de optimización en otras áreas de conocimiento e incluso en la gestión de carteras de valores[3], especialmente en los mercados accionarios y de divisas, donde existe mucha literatura al respecto. En cambio, es bastante importante el potencial para aplicar diferentes modelos de Machine Learning en carteras de renta fija, donde dichos modelos podrían ayudar a aumentar la precisión de las predicciones y mejorar el rendimiento de las carteras.

1.3. Objetivos

El propósito de este trabajo es utilizar algoritmos de Aprendizaje por Refuerzo para correr una estrategia, o un ensamble de estrategias, de gestión sobre un portafolio de renta fija, específicamente deuda soberana colombiana, y hacer una comparación de su desempeño, medido por la relación entre el rendimiento total del portafolio y el riesgo de mercado asumido, con otro tipo de metodologías.

Objetivos específicos

Este trabajo tiene como objetivos específicos los siguientes:

- Realizar un ciclo completo de ingeniería de datos, lo que implica recolectar los datos del mercado público de deuda soberana colombiana, realizar limpiezas generales y de outliers en caso de aplicar, hacer una descripción completa de las variables y los datasets, transformar los datos según las necesidades y características del mercado, basado en criterio experto profesional.
- Entrenar, probar y ensamblar modelos de Aprendizaje por Refuerzo para administrar una cartera de inversiones en el mercado de bonos soberanos de Colombia, utilizando para ello los datasets recolectados y las transformaciones a las que estos fueren

[3] (López de Prado, 2020) explica en su libro como el Machine Learning ha impactado positivamente la gestión de los asset managers, es decir, los gestores de carteras de inversión, toda vez que se minimizan tiempos de análisis de las potenciales inversiones, al tiempo que los mismos son, en general, más precisos. Todo esto ayuda al gestor a tomar decisiones más informadas, y, en general, mejores para su gestión específica.

sometidos, buscando alcanzar objetivos de rentabilidad y riesgo asumido, ajustando hiperparámetros según las necesidades.

- Evaluar el desempeño de los modelos, con base en diferentes métricas financieras y de ciencia de datos. Es importante en este objetivo tener en cuenta la necesidad de comparar el desempeño de las estrategias de Aprendizaje por Refuerzo con respecto a estrategias convencionales y frente a referencias -benchmarks- de mercado.

2. REVISIÓN DE LITERATURA

2.1 Aprendizaje de Máquina para Predicción del Mercado de Valores

El aprendizaje automático ha sido una solución común para los problemas financieros en los mercados de capitales (López de Prado, 2020), donde es ampliamente conocido, como explica Hull, que los precios de los instrumentos financieros que transan en mercados financieros siguen un movimiento browniano geométrico, que clasifica dentro del espectro de los procesos aleatorios markovianos (Hull, 2022), concepto que ha moldeado la forma en la que los agentes del mercado realizan valoraciones sobre instrumentos financieros derivados y de deuda desde mediados de los años setenta con la aparición de la teoría de valoración de opciones (Black & Scholes, 1973) y los aportes de Merton (1973) sobre la racionalidad de dichas valoraciones.

En ese orden de ideas, las aplicaciones del aprendizaje automático en los mercados financieros han sido un tema importante y siguen siendo una rama de investigación relevante para los mercados financieros que está en continuo desarrollo (Henrique y otros, 2019), en específico por la dificultad que se encuentra en la naturaleza aleatoria no estacionaria de las series de tiempo financieras (Zhang y otros, 2017). El inicio del aprendizaje automático en los mercados financieros está ligado con el auge de las redes neuronales a finales del siglo pasado (Refenes y otros, 1997), donde se trataron aplicaciones sobre predicción del mercado accionario con resultados iniciales satisfactorios mejores que resultados obtenidos por algunos análisis más tradicionales (Yoon y otros, 1993). Comenzando la siguiente década, otro tipo de algoritmos comenzaron a ser utilizados para la predicción del comportamiento de los mercados financieros, como Máquinas de Soporte Vectorial (Fernández-Rodríguez y otros, 2000).

Algunos precedentes a la estimación de nodos de la curva desde perspectivas de aprendizaje de máquina tuvieron que ver con un objetivo académico proveniente de la década de los ochenta, que pretendía establecer si una estrategia activa de administración de portafolios de renta fija en la que primase el tomar posiciones compradoras y vendedoras a lo largo de la curva de rendimientos era o no más rentable sobre una estrategia más conservadora en la que solo se comprasen y mantuviesen los títulos emitidos hasta el vencimiento. Inicialmente, Dyl y Joehnk (1981) concluyeron que esta estrategia fue más rentable que las

letras del tesoro americano, es decir, que las tasas de interés libres de riesgo de más corto plazo, entre los años 1970 y 1975. En otros estudios como Grieves y Marcus (1992) y Peláez (1997) se encontró evidencia empírica de que la estrategia activa fue superior al tradicional «comprar y mantener» en otros marcos temporales. En cambio, en Ang y otros (1998) y en Chua y otros (2005), la evidencia encontrada es mixta y no concluyente. Finalmente, Galvani y Landon (2013) sugieren que la estrategia activa es inefectiva desde un punto de vista de gestión de riesgos de mercado a través del concepto de mínima varianza cuando dicha estrategia incluye compras y ventas de títulos de largo plazo, situación atribuible a la mayor cantidad de convexidad y duración modificada que entonces recaería sobre el portafolio, incrementando el riesgo de mercado final (Fabozzi, The handbook of fixed income securities, 2021).

Dada la evidencia encontrada resumida en el párrafo anterior, otros autores comenzaron a explorar diversas técnicas de aprendizaje de máquina y su aplicación específica a la predicción de tasas de interés y de la curva de rendimientos, así como la optimización de estrategias activas sobre la curva, como, por ejemplo, la descrita en Zimmermann y otros (2000), en donde los autores encontraron que las técnicas convencionales para la predicción de diez nodos elegidos de la curva de rendimientos alemana son superados por una arquitectura de redes neuronales ajustadas por error de modelo, o en Gogas y otros (2015), quienes usaron variables macroeconómicas para modelar, con uso de Máquinas de Soporte Vectorial, la dirección de las tasas de interés y la ocurrencia de recesiones económicas, obteniendo resultados positivos en cuanto a la predicción de estos dos objetivos, además de superar modelos estadísticos convencionales estándar logit y probit.

2.2 Aprendizaje por Refuerzo: Teoría y Aplicación

El aprendizaje por refuerzo es un campo de estudio del aprendizaje de máquina, con características de aprendizaje no supervisado, y con un pasado que se remonta a los estudios de Bellman (1952) (1966), en donde el autor plantea los cimientos de lo que llamó «programación dinámica», cuyo principal propósito era crear algoritmos de optimización con capacidad de adaptarse a nuevos estados dentro de un espacio de posibilidades de esos

estados, haciendo un símil con la naturaleza humana del constante aprendizaje.

Este objetivo de optimización dentro de un espacio amplio de posibilidades de estados implica la necesidad de la existencia de un agente dentro del algoritmo con capacidad de llevar a cabo dicha optimización, lo que implica que los algoritmos de aprendizaje por refuerzo entreguen premios al agente por ejecutar una secuencia de decisiones basadas en probabilidad, que, correctas o incorrectas, lleven al agente a obtener la mayor cantidad de dicho premio. A esta secuencia probabilística de decisiones se le llama «política», mientras que el conjunto de estados posibles y de acciones, en conjunto, dada su aleatoriedad, permiten describir al aprendizaje por refuerzo como un Proceso de Decisión de Markov (Arulkumaran y otros, 2017).

La naturaleza aleatoria de los Procesos de Decisión de Markov -en adelante PDM-, representa un reto para el agente, puesto que se según lo mostrado en (Sutton, Temporal credit assignment in reinforcement learning, 1984), el número de acciones consecutivas que el agente puede ejecutar en cada marco temporal es limitado por las propias limitaciones existentes en el espacio de posibilidades. A manera de ejemplo, si en cierto PDM solo pueden darse un número determinado de estados diferentes, y, suponiendo que conocemos la combinación de acciones consecutivas que maximiza el premio para el agente, entonces cualquier acción incorrecta que el agente tome le impedirá alcanzar el premio óptimo en el futuro. Este problema para el agente se conoce como el problema de asignación temporal de crédito, y fue ampliamente abordado por Watkins (1989) en su tesis doctoral.

La solución de Watkins, bautizada como «Q-Learning», fue un hito que marcó una década de los noventa con valiosos aportes, estudios y variantes del aprendizaje por refuerzo. Dicha solución, basada en la ecuación de Bellman (1952), consiste en utilizar simulaciones basadas en métodos Monte Carlo para realizar un mapeo repetitivo y completo de todas las posibles políticas que el agente puede tomar dado cierto estado, así como sus potenciales premios, y utilizando un factor de descuento definido que permite darle más importancia a los premios de corto plazo (Arulkumaran y otros, 2017). Finalmente, el agente elige ejecutar la acción que tiene un potencial mayor premio. Posteriormente, en (Watkins & Dayan, 1992), se muestra como las decisiones tomadas por el agente convergen a las acciones óptimas cuando se realiza el mapeo completo de los estados y premios posibles al utilizar

Q-Learning.

Dentro del amplio espectro de trabajos sobre aprendizaje por refuerzo en los años noventa, se destacan estudios como (Sutton y otros, 1999), quienes encontraron que gracias al enfoque probabilístico de Q-Learning, es posible llegar a mayores niveles de abstracción en los problemas de aprendizaje por refuerzo no solo en PDM, sino también en PDM parciales, es decir, en donde no es posible para el agente observar todos los posibles estados dado un estado actual.

En la década siguiente surgieron de manera primitiva algunos de los algoritmos más utilizados de aprendizaje por refuerzo en la actualidad. En (Busoniu y otros, 2008) se cita una importante cantidad de estudios sobre aprendizaje reforzado multiagente, es decir, en cuyos algoritmos existe más de un agente capaz de ejecutar una política bajo diferentes incentivos o premios. También en esta década surgieron algoritmos y aplicaciones para aprendizaje por refuerzo con «gradiente de política», inspirados en el proceso de «backpropagation» para entrenamiento de redes neuronales (Kakade, 2001). Este tipo de algoritmos llegó a tener buenos resultados en aplicaciones relacionadas con la robótica, como por ejemplo en el trabajo de (Stone & Kohl, 2004), quienes utilizando este tipo de aprendizaje con gradiente de política como optimizador obtuvieron mejores resultados para entrenar a un robot cuadrúpedo a desplazarse, que con otro tipo de métodos. Otro tipo de algoritmos que también fueron ampliamente estudiados durante la época fueron los «actor-crítico», que consiste en incorporar otro agente cuyo objetivo es evaluar el desempeño del agente que busca maximizar el premio, con una clara inspiración en el surgimiento de las redes neuronales adversarias (Arulkumaran y otros, 2017). En (Konda & Tsitsiklis, 2003), los autores muestran como los «agentes críticos» ayudan a converger a los «agentes actores» a la solución óptima, guiándolos con su crítica hacia la dirección del gradiente dentro del espacio de posibles estados.

Con la llegada del aprendizaje profundo, el impacto en el aprendizaje por refuerzo fue significativo, igual que en otras ramas del aprendizaje de máquina. El incremento de la capacidad de procesar altas dimensionalidades por parte de redes neuronales cada vez más complejas, con más capas y diferentes tipos de funciones activadoras contribuyó, igualmente, a la experimentación y resolución de problemas más complejos, y con cada vez

mejor nivel de abstracción (Koutník y otros, 2013).

En el caso del aprendizaje por refuerzo, la llegada del aprendizaje profundo implicó una mejora sustancial de los algoritmos mencionados anteriormente, mientras que la popularización de la nube permitió incrementar la velocidad de cómputo e impulsar los trabajos con grandes volúmenes de información (LeCun y otros, 2015).

Trabajos como los de van Hasselt y otros (2015) y Gu y otros (2016) muestran como el uso de aprendizaje profundo mejora la convergencia de los algoritmos de Q-Learning tradicionales, además de una mejora importante en velocidad de cómputo que permite resolver problemas con mayor dimensionalidad, conociéndose así los nuevos algoritmos de Q-Learning profundo. De la misma manera sucede con otros algoritmos antes mencionados, mientras que al tiempo nuevos aportes, como el transfer-learning, entre otros, continúan acelerando la ola del aprendizaje profundo, y, en específico, el aprendizaje por refuerzo profundo (Wang y otros, 2022).

2.3 Aprendizaje por Refuerzo en Mercados Financieros

Los algoritmos Aprendizaje Reforzado comenzaron a ser utilizados para aplicaciones de negociación de los mercados financieros alrededor de la década de los noventa junto con toda la ola del aprendizaje de máquina de aquella época. Moody y Saffel (1998) muestran cómo un algoritmo Aprendizaje por Refuerzo Recurrente puede ser entrenado para el comercio de carteras de acciones, mientras que la optimización de la ratio de Sharpe, que es una medida de la relación de la rentabilidad total obtenida por asumir una cantidad de riesgo de mercado.

En la última década, con la aparición y auge del aprendizaje profundo, aparecieron también nuevos enfoques aprendizaje por refuerzo profundo (Sutton & Barto, 2018), con nuevas aplicaciones en diferentes mercados, desde el mercado mayorista energético en (Tao & Wencong, 2018), el mercado de divisas en (Carapuco y otros, 2018), o el mercado accionario, en donde los autores escribieron especialmente sobre aplicaciones bursátiles con diferentes enfoques -como de Q-Learning profundo en (Carta y otros, 2021)-, y estrategias

adaptativas de negociación de acciones fueron el centro estudios como (Wu y otros, 2020). En todos los casos el rendimiento mejoró con respecto a los métodos de aprendizaje por refuerzo y redes neuronales básicas y los métodos convencionales de comprar y mantener.

Yang y otros (2020), específicamente propusieron un conjunto de diferentes algoritmos de aprendizaje por refuerzo basados en la arquitectura «actor-crítico», como Advantage Actor Critic A2C, Proximal Policy Optimization PPO y Deep Deterministic Policy Gradient DDPG, para la negociación de acciones y la administración de una cartera de renta variable. La estrategia ensamblada funcionó y obtuvo buenos resultados incluso durante la crisis del covid-19.

Por otra parte, el aprendizaje automático para aplicaciones de renta fija nunca se ha centrado en el comercio o la gestión y optimización de carteras (Dixon y otros, 2020). La atención se centró en el modelado de la curva de rendimiento, la predicción de su forma, sus movimientos y, en algunos casos, en predicción de la próxima crisis financiera, utilizando máquinas de soporte vectorial -SVM- y análisis de componentes principales -PCA-, entre otros métodos (Gogas y otros, 2015). El aprendizaje por refuerzo no ha sido estudiado en profundidad bajo la óptica de la Renta Fija, por ello, en su tesis doctoral Nunes (2022) realiza un diagnóstico de dicho vacío en la literatura y se dispone a proponer diferentes algoritmos, encontrando que los algoritmos DDPG tenían un mejor rendimiento en la negociación de ETFs de renta fija. Sin embargo, los ETF, pese a poder tener como subyacente uno o varios instrumentos de renta fija, pueden entenderse como instrumentos de renta variable, por lo que valdría la pena revisar si estos algoritmos de aprendizaje por refuerzo profundo pueden utilizarse directamente sobre los subyacentes de renta fija.

A modo de comentario de final, se resalta que dentro de la revisión de literatura realizada no se encontraron precedentes científicos ni evidencia de aproximaciones empíricas de aplicaciones puntuales de aprendizaje por refuerzo a los mercados de renta fija colombiana, ni a la gestión activa de portafolios con títulos de ese tipo.

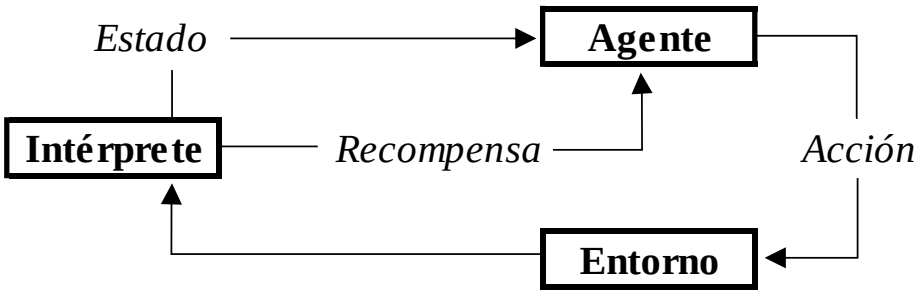
3. MARCO TEÓRICO

3.1 Aprendizaje por Refuerzo

El Aprendizaje por Refuerzo es una rama del aprendizaje automático donde un agente inteligente aprende cómo actuar dentro de un entorno, buscando maximizar las recompensas a largo plazo dadas por un intérprete en función de ciertos objetivos de rendimiento definidos previamente (Wang y otros, 2022).

En la Figura 1 se muestra un esquema general de cómo funciona el aprendizaje por refuerzo. En la primera iteración, el agente inteligente 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) (Sutton & Barto, 2018). Los resultados de esa 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 anteriores y aplicará ese conocimiento en futuras iteraciones dentro del entorno, buscando maximizar la cantidad de recompensa que recibe del intérprete.

Figura 1. Diagrama de flujo típico de un algoritmo de aprendizaje por refuerzo



Los primeros algoritmos de RL fueron entrenados para resolver problemas en entornos de baja dimensión (Sutton & Barto, 2018). Sin embargo, con los años aparecieron problemas de mayor envergadura, y con la aparición de las redes neuronales profundas, los algoritmos RL comienzan a ser más complejos, eficientes y útiles para resolver los problemas más

complejos y de mayor envergadura, dando cabida a los algoritmos de Aprendizaje por Refuerzo Profundo (Arulkumaran y otros, 2017).

Los siguientes conceptos son importantes para una buena comprensión del aprendizaje por refuerzo:

3.1.1 Procesos de Decisión de Markov (PDM):

Es un marco común para resolver problemas de aprendizaje por refuerzo que consiste en algunos supuestos, como, por ejemplo, que el entorno es markoviano y observable (Sutton & Barto, 2018) o parcialmente observable (Sutton y otros, 1999). 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 (Yang y otros, 2020) (Wang y otros, 2022):

- Un estado s en el que se encuentra el agente, y que pertenece a un set de posibles estados S. El estado inicial es s_0.
- Una acción a que el agente toma en 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'.
- 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 (Watkins C. J., 1989).
- 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.

3.1.2 Ecuación de Bellman y Q-Learning:

Dado el número de trayectorias, la política y los diferentes estados a los que se puede enfrentar el agente dentro del entorno, entonces es necesario calcular la recompensa esperada del agente por encontrarse en cierto estado, por lo que Bellman (1966) propone una Función de Valor del Estado $\mathbf{V}(\mathbf{s})$**,** mejorada en (Sutton, 1984) que a través de recursividad permite encontrar una valor de recompensa para el estado actual teniendo en cuenta los posibles estados futuros traídos a valor presente con el factor de descuento $\boldsymbol{\gamma}$ planteado, tal como se muestra en la ecuación (1).

$$V(s) = E_{\pi}[\rho_{t+1} + \gamma V(s_{t+1}) | s_t = s]. \quad (1)$$

En la misma línea, Watkins (1989) 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 $\boldsymbol{a}$ partiendo de un estado $\boldsymbol{s}$ es la Función de Valor Estado-Acción $\mathbf{Q}(\mathbf{a}, \mathbf{s})$, ecuación (2) (Watkins & Dayan, 1992).

$$Q(a, s) = E_{\pi}[\rho_{t+1} + \gamma Q(a_{t+1}, s_{t+1}) | a_t = a,\ s_t = s]. \quad (2)$$

La Función de Valor Estado-Acción ayuda a mapear acciones futuras ligadas a un estado futuro relacionado, permitiendo así una convergencia más rápida y precisa en cantidad de instancias, pero un poco más exigente respecto a la capacidad de cómputo (Sutton & Barto, 2018).

3.1.3 Métodos de Aprendizaje por Refuerzo Profundo:

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 (Arulkumaran y otros, 2017), y otras soluciones comenzaron a ser planteadas con el objetivo de lograr eficiencias necesarias en este tipo de entornos de grandes dimensiones, la gran mayoría siendo PDM parcialmente observables. En el desarrollo de este trabajo se

utilizarán tres métodos específicos a describir a continuación y métodos de ensamble que involucran estos métodos.

- Métodos de Actor-Crítico

  Estos métodos son la 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 optimizan utilizando la política, pero teniendo en cuenta la función de valor implícita crítico

  Como los métodos Actor-Crítico típicos (Konda & Tsitsiklis, 1999), A2C utiliza un agente adicional, el crítico, que se encarga de revisar si la política seguida por el agente que la ejecuta, el actor. Ambos agentes actúan de manera independiente.

  Particularmente, un algoritmo de A2C entrega al agente crítico el resultado de la función de ventaja, ecuación (3), dada por la diferencia entre la función de Valor Estado-Acción, y la función de Valor del Estado (Yang y otros, 2020), lo que se interpreta como la ventaja de usar una política $\boldsymbol{\pi}$ sobre la Función de Valor del Estado por si sola.

$$A(a,s) = Q(a,s) - V(s) \tag{3}$$

- Métodos de Gradiente de Política Determinístico:

  Los algoritmos de gradiente de política son ampliamente utilizados en aprendizaje por refuerzo para problemas en donde las acciones que puede tomar el agente actor se pueden describir con una función de probabilidad continua. A partir de allí, estos algoritmos mapean la función $Q(a,s)$, y a través de la optimización con gradiente, buscan minimizar el error total o maximizar la recompensa esperada (Silver y otros, 2014).

  DDPG, es un algoritmo de gradiente de política que no utiliza funciones de probabilidad sobre variables continuas. En cambio, también permite optimizar el gradiente de política, ecuación (4), teniendo en cuenta un número de acciones posibles finitas y determinísticas para una función objetivo de retorno J, ecuación (5) (Li y otros, 2019), lo que representa una ventaja, específicamente, para los problemas

relacionados con trading, debido a que las acciones que el agente actor puede tomar cumplen esa condición determinística: Compra, venta, mantener la posición, entre otras que puedan plantearse (Yang y otros, 2020).

- Métodos Basados en Regiones de Confianza:

  Mientras se entrena un algoritmo Aprendizaje automático, podrían existir iteraciones de diferente longitud para los conjuntos de datos de entrenamiento en los que la política de cada iteración podría ser diferente de otra iteración. La longitud óptima de las iteraciones es aquella en la que las políticas convergen o tienen bastante similitud entre una iteración y otra. Los métodos basados en regiones de confianza buscan mejorar la estabilidad del entrenamiento entre pasos, de forma que las políticas entre un paso y otro no diverjan mucho (Schulman y otros, 2015).

  Los algoritmos PPO, descritos en (Schulman y otros, 2017) son algoritmos de gradiente de política y basados en regiones de confianza que buscan mantener controlada la actualización de la política a través del tiempo, es decir, en cada iteración, de manera que una nueva política en un estado específico no sea sustancialmente diferente a la política de la iteración pasada, logrando una mejor estabilidad en la función objetivo. Para ello, se introduce dentro de la función objetivo, ecuación (6), una relación entre la política nueva y la antigua, ecuación (7), que ponderará la función de ventaja estimada para la nueva política.

$$\nabla_\theta J(\theta) = E_s[\nabla_\theta \pi_\theta(s) \nabla_a Q_\pi(a,s) | a = \pi_\theta(s)\,] \tag{4}$$

$$J(\theta) = E_s[Q_\pi(a,s)\,] \tag{5}$$

$$J(\theta) = E_s\big[\min\left(r(\theta)\hat{A}_\pi(a,s), recorte(r(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_\pi(a,s)\right)\big] \tag{6}$$

$$r(\theta) = \frac{\pi_\theta(a,s)}{\pi_{\theta_{antigua}}(a,s)} \tag{7}$$

El término $recorte(r(\theta), 1-\varepsilon, 1+\varepsilon)$, recorta la razón entre las políticas actual y antigua para que esta no pueda moverse por fuera del intervalo $[1-\varepsilon, 1+\varepsilon]$, asegurando así que la política sea relativamente similar a la política antigua. El valor de ε es considerado un parámetro importante para que un modelo de PPO escoja

políticas más o menos similares a las anteriores; con un mayor ε se aceptarán políticas más distantes y menos estables entre sí, mientras que un valor menor de ε no las permitirá, y por tanto las nuevas políticas serán más próximas a la política anterior, ayudando a mantener una estabilidad en el tiempo.

3.2 Gestión de Portafolios de Renta Fija

Los portafolios de renta fija juegan un importante papel en la gestión del balance de activos y pasivos, y en gestión patrimonial, tanto para instituciones financieras como para agentes independientes en los mercados financieros. Estos portafolios están compuestos por instrumentos de renta fija, los cuales son, en esencia, deuda emitida por una contraparte (Fabozzi, The handbook of fixed income securities, 2021).

A continuación, presentamos una serie de conceptos necesarios para el buen entendimiento del funcionamiento de este tipo de portafolios:

3.2.1. Instrumentos de Renta Fija

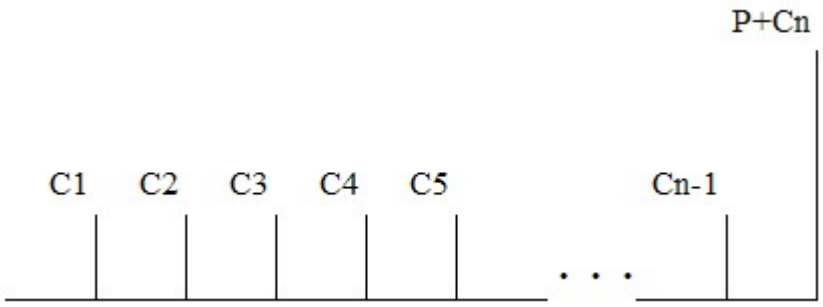
Un instrumento de renta fija es aquel en el que el inversor compra el derecho a los flujos de caja futuros que el emisor o prestatario ha prometido pagar. Este derecho puede negociarse en un mercado secundario, y el precio justo vendrá determinado por el rendimiento al vencimiento[4], que es la tasa de descuento exigida por los compradores que incorpora todas las expectativas y riesgos implícitos sobre el valor (Fabozzi, The handbook of fixed income securities, 2021). La figura 2 muestra el flujo de caja de un instrumento de renta fija convencional, con cupones periódicos y el pago del principal nocional al final, con el último cupón.

En este trabajo, se estudian carteras constituidas con bonos soberanos convencionales emitidos por el gobierno colombiano conocidos como Títulos de Tesorería Clase B, o

[4] Rendimiento al Vencimiento, en inglés, Yield To Maturity (YTM), es la tasa de descuento justa a la que se descuentan los flujos de caja futuros del título negociado para que el precio de este sea considerado por el mercado como el «precio justo». Esto implica que ese precio justo es igual a los flujos de caja descontados, y, por tanto, si el inversionista compra el título y decir mantenerlo hasta su fecha de vencimiento, entonces su Tasa Interna de Retorno (TIR), será igual a la YTM.

simplemente, TES Clase B. Sin embargo, otros tipos de bonos y emisores pueden ser incluidos en carteras de renta fija.

Figura 2. Diagrama de flujo de caja de un instrumento de renta fija convencional



En ese orden de ideas, la ecuación (8) expresa el precio limpio (PL) de un TES Clase B en términos de los flujos de caja futuros (CF) descontados por el rendimiento al vencimiento (YTM), que es al mismo tiempo la tasa de negociación del mercado. Podemos reexpresar el Precio Limpio en la ecuación (9) si tomamos cada cupón (C) del bono como el resultado de la multiplicación entre el nocional o principal (P) del bono en base 100 o base 1 por la tasa cupón ofrecida por el emisor en las condiciones faciales del título, y, si sustituimos los cupones periódicos por la fórmula del valor presente de anualidad (Fabozzi, The handbook of fixed income securities, 2021), siendo ***n*** la cantidad de anualidades o cupones totales entregados durante el flujo de caja restante.

$$PL = \sum_{i=1}^{n} \frac{\mathrm{CF}_i}{(1+YTM)^i} \tag{8}$$

$$PL = \frac{C}{YTM}\left[1 - \frac{1}{(1+YTM)^n}\right] + \frac{P=1}{(1+YTM)^n} \tag{9}$$

$$PS = PL(1+YTM)^{d/b} \tag{10}$$

Dado que el PL asume que la negociación se da exactamente en la fecha de emisión o en la fecha de pago de un cupón, se hace necesario recalcular el precio justo de la transacción cuando esta se encuentra por fuera de las fechas anteriormente descritas. Esto es importante porque para esas fechas no exactas, el PL no toma en cuenta el valor de los intereses que se han causado sobre el siguiente cupón, y que, teóricamente, le pertenecen a la parte vendedora en la transacción dado que ha mantenido el título durante esos días adicionales a la fecha de

pago del último cupón. Es así como en la ecuación (10) se calcula el Precio Sucio (PS) del título como el valor futuro del PL a los ***d*** días transcurridos desde el pago del último cupón a la fecha de negociación, y tomando como base una cantidad de ***b*** días, que equivale a la cantidad de días que hay entre el último cupón pagado y el siguiente por pagar según el flujo de caja definido para ese instrumento.

Cuando un agente compra un bono en el mercado secundario pagando el PS por él, y mantiene ese título hasta su fecha de vencimiento o de maduración, entonces su rentabilidad neta será exactamente el YTM de mercado asociado a ese PS que pagó, y, en consecuencia, no existe riesgo de mercado asociado al periodo de tenencia para dicho inversionista.

En cambio, el riesgo de mercado para el inversionista puede surgir si decide no mantener su inversión en ese título hasta la fecha de vencimiento de este, es decir, que planea venderlo antes de su maduración. En ese caso, para el momento de la venta, las condiciones del mercado, y por tanto la YTM con la que se descuentan los flujos de caja del título, pueden haber cambiado, generando cambios en la valoración del título, y así, posibilidad de ganar o perder dinero por cuenta de una variación en los precios de mercado (Fabozzi, The handbook of fixed income securities, 2021).

Ahora bien, en el marco de la administración de activos y pasivos, y de la gestión patrimonial, realizar un estricto control al riesgo de mercado resulta fundamental. Previo al control y administración del riesgo de mercado, los gestores de portafolio primero lo deben medir, y para ello, exploraremos los siguientes conceptos (Fabozzi, 2007):

- Duración de Macaulay, ecuación (11):

  La Duración Macaulay se define como la media ponderada de tiempo de los flujos de caja de un título dado el Precio Sucio actual. En otras palabras, puede interpretarse como el tiempo medio en el que se recupera la inversión, dado el Precio Sucio pagado.

$$DMac = \frac{n}{PS}\sum_{i=1}^{n}\frac{\mathrm{CF}_i}{(1+YTM)^i} \tag{11}$$

- Duración Modificada, ecuación (12):

Como se mencionaba antes, el riesgo de mercado en las operaciones de renta fija está asociado a las variaciones de los precios de los títulos debidas a las variaciones de los tipos de interés, ya que los primeros reaccionan, en cierta medida, de forma inversa a los segundos. La sensibilidad de la variación del precio respecto a la variación del tipo de interés se conoce como Duración Modificada. Así, un cambio de cien puntos básicos (pbs) en la YTM de mercado implica un cambio porcentual en el Precio Sucio equivalente a la Duración Modificada. Matemáticamente, se puede entender a la Duración Modificada como la primera derivada del Precio Sucio con respecto a la YTM.

$$DMod = \frac{DMac}{1+YTM} \tag{12}$$

- Convexidad, ecuación (13):

Como se muestra en la figura 3, cuando se presentan grandes cambios en el tipo de interés, la duración modificada no es suficiente para describir la sensibilidad sobre el precio. Esto es porque la relación entre el Precio Sucio y la YTM no es lineal, y, en lugar de ello, es convexa e inversa, por lo que cada bono tiene una convexidad asociada a su precio.

Matemáticamente, la convexidad de un título es la segunda derivada del Precio Sucio con respecto al YTM. Se puede interpretar como una medida geométrica de la volatilidad del bono. Una convexidad mayor implica mayores incrementos del Precio Sucio causados por una caída de la YTM y menores caídas del Precio Sucio causadas por el mismo incremento absoluto de la YTM.

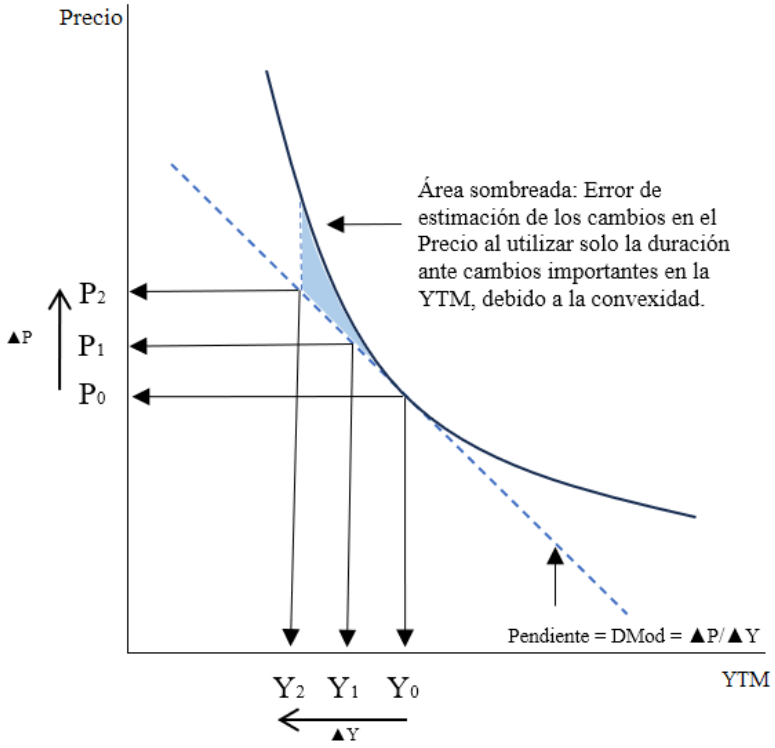
$$CVX = \frac{n}{PS}\sum_{i=1}^{n}\frac{\mathrm{CF}_i}{(1+YTM)^i} \tag{13}$$

$$\frac{\Delta PS}{PS} = (-DMod\ \times \Delta YTM) + (\frac{1}{2}\times CVX \times \Delta YTM^2) \tag{14}$$

Al combinar los conceptos de duración y convexidad de un instrumento de renta fija, entonces ya es posible calcular de manera precisa el cambio en el precio sucio de un bono cuando se presentan cambios significativos en la tasa de interés utilizando la

ecuación (14). Nótese que, el segundo término de la ecuación, al elevarse al cuadrado, es aproximadamente igual a cero cuando los cambios en la YTM son muy pequeños.

Figura 3. Relación entre la YTM de un bono y su precio. Se evidencia convexidad en el precio que la duración modificada por sí sola no puede estimar.



- Dollar Duration o DV01, ecuación (15):

La duración y convexidad de un título dependen de múltiples factores, como la tasa y la frecuencia de los cupones, la YTM y el plazo al vencimiento del título. El riesgo de mercado no es comparable entre dos bonos diferentes y con diferencias en esos factores, y ahí radica la importancia de la DV01, o Dollar Duration, que permite comparar el riesgo de mercado dos títulos diferentes, midiendolo en términos monetarios. El DV01 es la ganancia o pérdida en una posición en un bono o un portafolio de bonos debido a un incremento de un punto básico en la YTM. Reducir el DV01 en una cartera es una de las actividades más importantes en la gestión de carteras de renta fija. El DV01 puede reducirse vendiendo en corto otros bonos.

$$DV01 = -DMod \times PL \times 0.01\% \quad (15)$$

3.2.2 La Curva de Rendimientos

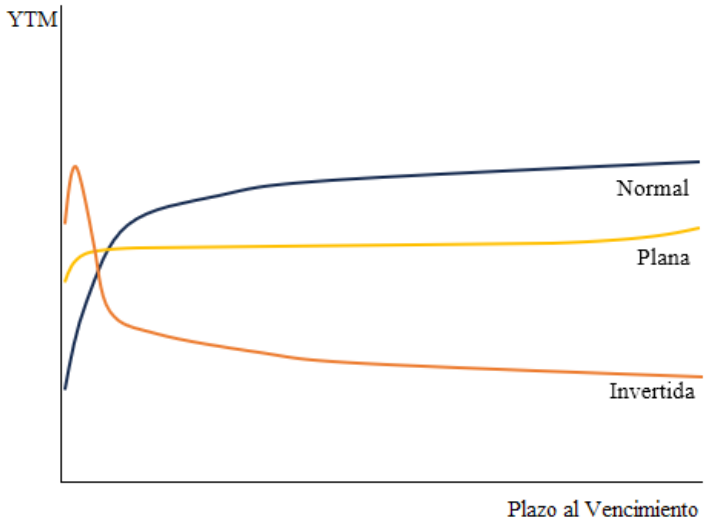
La curva de rendimientos es una representación gráfica de la relación entre los vencimientos y los rendimientos de los bonos dentro de un mercado específico (Fabozzi, 2007). Una representación de la curva de rendimientos incluye títulos con las siguientes características:

- Con la misma indexación
- Con la misma denominación monetaria
- Con la misma calificación crediticia
- Con diferentes vencimientos

Normalmente, la pendiente de la curva de rendimientos es positiva, sin embargo, dada una situación económica concreta, la curva podría adoptar otra forma. A continuación, se indican las formas que podría adoptar la curva de rendimientos (Figura 4):

- *Normal:* Las tasas de interés a corto plazo son más bajos que a largo plazo.
- *Plana:* Las tasas de interés plazo son similares a las de largo plazo.
- *Invertido:* Las tasas de interés a corto plazo son superiores a las de largo plazo.

Figura 4. Formas de la curva de rendimientos.



Hay una variedad de factores económicos que influyen en la forma de la curva de rendimiento, y hay diferentes teorías sobre cómo esos factores explican la forma de la curva y sus movimientos (Pelaez, 1997). Sea cual sea la explicación de los movimientos de la curva, los operadores se aprovechan de ellos comprando o vendiendo títulos en diferentes nodos de la curva. En términos generales, existen tres movimientos básicos para la curva de rendimientos:

- *Desplazamiento Paralelo:* El diferencial entre las tasas de interés a largo y corto plazo no cambia. Las tasas aumentan o disminuyen en la misma cantidad de puntos básicos a lo largo de la curva.

- *Empinamiento:* El diferencial entre las tasas de interés a largo y a corto plazo aumenta. Ocurre cuando las tasas de interés a corto plazo bajan mientras que las tasas de interés a largo plazo suben, o cuando las tasas de interés a corto plazo suben menos que el incremento de las tasas de interés a largo plazo.

- *Aplanamiento:* El diferencial entre las tasas de interés a largo y a corto plazo disminuye. Ocurre cuando las tasas de interés a corto plazo aumentan mientras que las tasas de interés a largo plazo bajan, o cuando las tasas de interés a corto plazo bajan menos que el incremento de las tasas de interés a largo plazo.

Los operadores estructuran sus estrategias y carteras de renta fija en función de las expectativas sobre los movimientos de la curva. Por ejemplo, cuando un inversor espera un aplanamiento, probablemente financiará compras de bonos a largo plazo con ventas en corto de bonos a corto plazo. Si espera lo contrario, es decir, un empinamiento, entonces financiará compras de bonos a corto plazo con ventas en corto de bonos a largo plazo. La proporción entre las posiciones cortas y largas será la necesaria para que el DV01 del portafolio completo sea igual o cercano a cero, es decir, se trata de un problema de minimización del riesgo de mercado en términos monetarios.

4. MODELACIÓN

4.1 Descripción y Exploración de los Datos

4.1.1 Recolección de los Datos

Los datos recopilados corresponden a los cierres individuales de operaciones de compraventa o simultáneas registrados en el mercado de renta fija soberana de títulos de tesorería (TES) entre enero de 2015 y diciembre de 2022 a través del SEN (Sistema Electrónico de Negociación), administrado por el Banco de la República (2022).

4.1.2 Procesamiento y Estructuración de los Datos

Cada uno de los registros, corresponde a una transacción pactada entre dos contrapartes, de las cuales solo se tomaron en cuenta las operaciones en la rueda de contado, en la cual se encuentran las compraventas efectivas entre las dos contrapartes sobre TES a un precio determinado por ellas, con cumplimiento el mismo día. No se tienen en cuenta las operaciones de la rueda de simultáneas por ser estas operaciones de liquidez, es decir, préstamos entre agentes del mercado garantizados con TES, por lo cual estas operaciones no aportan a la formación de precios del mercado. Las transacciones son agrupadas por título y por día, ponderando la YTM negociada por los montos individuales de giro de cada una de las transacciones intradías de cada título.

Es muy importante tener en cuenta que cuando un título no se negocia durante un día bursátil, entonces este se valora utilizando la última YTM negociada, por lo que este será el método para rellenar datos faltantes, teniendo en cuenta, claro, si el título ya fue emitido y no ha madurado. Con esta consideración, el set de datos inicial, cuya estructura se evidencia en la tabla 1, queda con cinco columnas, así:

- Fecha: Fecha a la cual un título tiene una tasa cualquiera promedio ponderada.

- Instrumento: Se refiere al nemotécnico del título negociado.[5]

[5] El nemotécnico de un TES tiene la siguiente estructura según la posición de sus caracteres:
Ejemplo TFIT16240724.
• 1: TES Clase B.
• 2-4: Tipo de Tasa y Amortización
• 5-6: Años comprendidos entre la emisión y el vencimiento

- Tasa: Se refiere la tasa promedio ponderada negociada para un título en cierta fecha.

- Fecha_Ems: Es la fecha de emisión del título.

- Fecha_Vto: Es la fecha de vencimiento del título.

Tabla 1. Primeros dos días del set de datos inicial, limpio de datos faltantes y que será utilizado para obtener campos calculados necesarios para el análisis.

Fecha	Instrumento	Tasa	Fecha_Ems	Fecha_Vto
5/01/2015	TFIT16240724	7.123	24/07/2008	24/07/2024
5/01/2015	TFIT10040522	6.902	4/05/2012	4/05/2022
5/01/2015	TFIT15260826	7.32	26/08/2011	26/08/2026
5/01/2015	TFIT07150616	5.044	15/06/2009	15/06/2016
5/01/2015	TFIT06211118	5.873	21/11/2012	21/11/2018
5/01/2015	TFIT06110919	5.944	11/09/2013	11/09/2019
5/01/2015	TFIT15240720	6.266	24/07/2005	24/07/2020
5/01/2015	TFIT16280428	7.618	28/05/2012	28/04/2028
6/01/2015	TFIT16240724	7.172	24/07/2008	24/07/2024
6/01/2015	TFIT10040522	6.944	4/05/2012	4/05/2022
6/01/2015	TFIT15260826	7.343	26/08/2011	26/08/2026
6/01/2015	TFIT07150616	5.062	15/06/2009	15/06/2016
6/01/2015	TFIT06211118	5.867	21/11/2012	21/11/2018
6/01/2015	TFIT06110919	5.958	11/09/2013	11/09/2019
6/01/2015	TFIT16280428	7.674	28/05/2012	28/04/2028
6/01/2015	TFIT15240720	6.266	24/07/2005	24/07/2020

Los datos faciales de los títulos se utilizan posteriormente para agregar las siguientes columnas al set de datos, así:

- Días_vto: Son los días al vencimiento del título negociado y se calculan como la diferencia en días entre las columnas Fecha_Vto y Fecha.

- PL: Precio Limpio del título negociado en base 1, usando la ecuación (9).

- PS: Precio Limpio del título negociado en base 1, usando la ecuación (10).

- Dmac: Duración de Macaulay del título negociado, usando la ecuación (11).

- Dmac: Duración Modificada del título negociado, usando la ecuación (12).

• 7-12: Fecha de vencimiento en formato DDMMA

- DV01: Dollar Duration en pesos colombianos del título negociado, usando la ecuación (15).

- CVX: Convexidad del título negociado, usando la ecuación (13).

La estructura final de los datos se muestra en la tabla 2, en donde se muestran las columnas del set de datos inicial más las columnas con campos calculados. Es muy importante tener en cuenta que cada uno de los títulos tiene un periodo de vigencia diferente, esto es, el periodo entre su emisión y vencimiento, por lo que durante el periodo de análisis se presenta que el número de títulos diferentes varíe, dado que algunos de ellos ya vencieron o no han sido emitidos, esto es, que las series de tiempo de cada título no necesariamente tienen los siete años completos.

Tabla 2. Primeros dos días del set de datos final, limpio de datos faltantes y que será utilizado por los modelos para realizar operaciones de compraventa.

Fecha	Instrumento	Tasa	PL	PS	Dmac	Dmod	DV01	CVX	Fecha_Ems	Fecha_Vto	Dias_vto
5/01/2015	TFIT16240724	7.123	1.201	1.238	6.608	6.169	-$ 616,907	60.914	24/07/2008	24/07/2024	3,488
5/01/2015	TFIT10040522	6.902	1.006	1.052	5.725	5.355	-$ 535,525	44.057	4/05/2012	4/05/2022	2,676
5/01/2015	TFIT15260826	7.32	1.014	1.040	7.986	7.442	-$ 744,175	87.746	26/08/2011	26/08/2026	4,251
5/01/2015	TFIT07150616	5.044	1.041	1.070	1.377	1.311	-$ 131,121	3.328	15/06/2009	15/06/2016	527
5/01/2015	TFIT06211118	5.873	0.970	0.976	3.598	3.398	-$ 339,816	17.070	21/11/2012	21/11/2018	1,416
5/01/2015	TFIT06110919	5.944	1.045	1.064	4.087	3.857	-$ 385,740	22.181	11/09/2013	11/09/2019	1,710
5/01/2015	TFIT15240720	6.266	1.231	1.265	4.385	4.127	-$ 412,681	26.543	24/07/2005	24/07/2020	2,027
5/01/2015	TFIT16280428	7.618	0.869	0.909	8.490	7.889	-$ 788,919	99.609	28/05/2012	28/04/2028	4,862
6/01/2015	TFIT16240724	7.172	1.197	1.235	6.601	6.159	-$ 615,902	60.810	24/07/2008	24/07/2024	3,487
6/01/2015	TFIT10040522	6.944	1.003	1.050	5.720	5.348	-$ 534,847	43.999	4/05/2012	4/05/2022	2,675
6/01/2015	TFIT15260826	7.343	1.012	1.038	7.980	7.434	-$ 743,432	87.646	26/08/2011	26/08/2026	4,250
6/01/2015	TFIT07150616	5.062	1.041	1.070	1.375	1.308	-$ 130,836	3.318	15/06/2009	15/06/2016	526
6/01/2015	TFIT06211118	5.867	0.970	0.977	3.595	3.396	-$ 339,579	17.048	21/11/2012	21/11/2018	1,415
6/01/2015	TFIT06110919	5.958	1.044	1.063	4.084	3.854	-$ 385,412	22.154	11/09/2013	11/09/2019	1,709
6/01/2015	TFIT16280428	7.674	0.865	0.905	8.477	7.873	-$ 787,306	99.396	28/05/2012	28/04/2028	4,861
6/01/2015	TFIT15240720	6.266	1.231	1.265	4.383	4.124	-$ 412,424	26.517	24/07/2005	24/07/2020	2,026

4.1.3 Exploración de los Datos

Dada la particularidad de este set de datos de tener series de tiempo con diferentes periodos de vigencia cada una, se propende por realizar análisis puntuales sobre los datos, divididos en dos grupos: Análisis sobre el comportamiento de las variables en la serie de tiempo de un solo título, y análisis sobre los movimientos de la curva de rendimientos.

- Análisis individual de títulos: Tomando como referencia el TES tasa fija con vencimiento en abril de 2028, cuyo nemotécnico es TFIT16280428, realizaremos un

análisis sobre el comportamiento de las variables del set de datos a lo largo del tiempo. Este título tiene la particularidad de encontrarse vigente durante todo el periodo de la muestra, los siete años completos, puesto que fue emitido en 2012 para tener una vida de dieciséis años en total, y así vencer en 2028.

En la figura 5, inicialmente, se muestra la relación que tienen el precio limpio (rojo, eje izquierdo), el precio sucio (azul, eje izquierdo) y la YTM (verde, eje derecho), evidenciándose que los precios sucio y limpio mantienen una senda de movimiento similar, y cuya diferencia radica en el valor de los intereses acumulados pendientes de pago; en cambio, se presenta una clara correlación inversa en entre el precio y la tasa de interés, lo que es consecuencia directa de la formación del precio limpios con la ecuación (8), ello sin llegar a ser una relación lineal debido a la existencia de la convexidad, como se muestra en la figura 3. La tabla de correlaciones entre las variables, tabla 3, muestra las relaciones anteriormente descritas.

Dada la alta correlación entre precio sucio y precio limpio, para los análisis posteriores en esta sección de exploración de datos, solo será tomado como precio el precio limpio, pues este refleja mejor el comportamiento de los precios de mercado en términos de volatilidad, puesto que no incluye el efecto de los intereses acumulados no pagados, o carry (Fabozzi, 2007).

Figura 5. TFIT16280428: Precio Sucio, Precio Limpio y YTM. 2015-2022. Precio limpio (rojo), precio sucio (azul) en eje izquierdo, YTM (verde) en eje derecho.

Tabla 3. Correlación PS, PL y YTM para TFIT16280428. 2015-2022

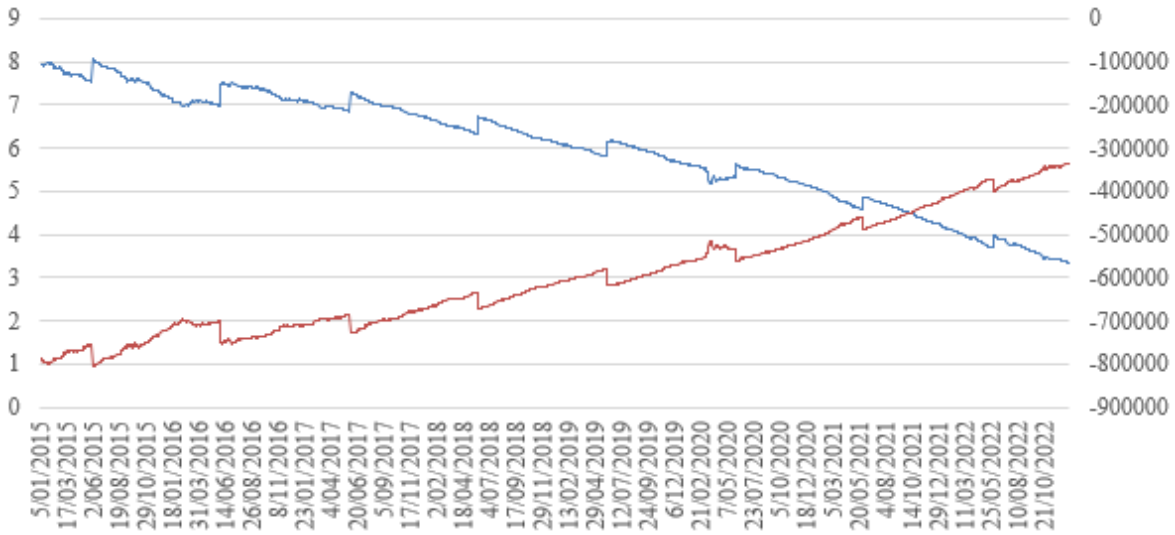
	Precio Sucio	*Precio Limpio*	*YTM*
Precio Sucio	1.00	0.97	-0.90
Precio Limpio	0.97	1.00	-0.92
YTM	-0.90	-0.92	1.00

Por otro lado, la duración de Macaulay y la duración modificada son dos variables que, en general, dependen más del tiempo restante al vencimiento que de la formación de precios en el mercado de valores, como se evidencia en la tabla 4, en donde la correlación de las duraciones con la tasa de rendimiento es muy baja, entendiéndose que no puede siquiera asumirse algún tipo de relación entre las variables, y por tanto, tampoco entre las duraciones y el precio, dado que el precio es considerado como un accidente de la YTM.

Tabla 4. Correlación DMac, DMod y YTM para TFIT16280428. 2015-2022

	Dmac	*Dmod*	*YTM*
Dmac	1.00	1.00	-0.29
Dmod	1.00	1.00	-0.34
YTM	-0.29	-0.34	1.00

Figura 6. TFIT16280428: Duración Modificada Vs DV01. 2015-2022. Duración modificada (azul) en eje izquierdo, DV01 (rojo) eje izquierdo.



Por último, en términos de riesgo, y dado que la duración (azul, eje izquierdo) termina dependiendo principalmente del plazo al vencimiento, se encuentra que, según la figura 6, esta es inversa al valor del DV01 (rojo, eje derecho) en pesos para una posición direccional larga de mil millones de pesos en valor de giro, un supuesto ampliamente aceptado por los operadores de renta fija, esto es, que los títulos con vencimientos más lejanos son, en general, más propensos al riesgo de mercado que los títulos con vencimientos más cortos.

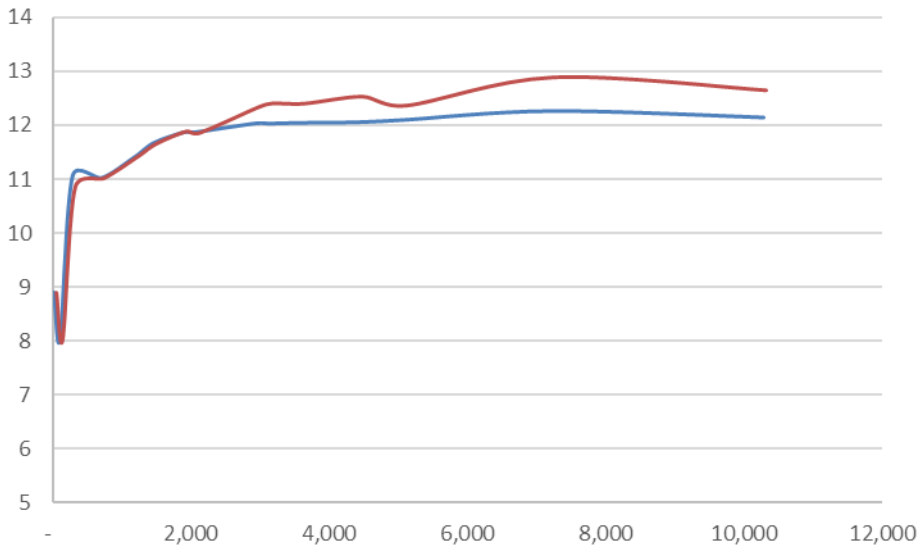
- Análisis de movimientos de la curva: Como bien se menciona en la sección 3.2 – B de este trabajo, los títulos según su plazo pueden tener comportamientos diferentes según las expectativas existentes sobre varios factores económicos y financieros para diferentes plazos, lo que hace que los retornos de los títulos, al graficarse en función del plazo al vencimiento conformen una curva de rendimientos que puede tener, en general, cualquiera de las formas de la figura 4.

  Si las condiciones que le daban la forma a la curva de rendimientos en un momento determinado cambian para otro momento, entonces la forma de la curva puede cambiar, presentando un desplazamiento paralelo, un empinamiento o un aplanamiento. A continuación, se describen las relaciones que tuvieron los títulos del set de datos en cuanto a sus precios y tasas de mercado cuando se presentaron movimientos curva durante un aplanamiento (ver figura 7).

Ese aplanamiento de la curva ocurrido en el mes de agosto de 2022, encuentra a la curva de rendimientos preponderantemente por encima del 10% en su YTM. El aplanamiento se da porque, pese a que las tasas de interés de corto plazo se mantienen inalteradas, para el mediano y largo plazo las tasas de interés tuvieron una caída. Los movimientos de las tasas para tres títulos TES específicos, uno de corto, otro de mediano y el último de largo plazo, se observan en la figura 8, observándose que el spread o diferencial entre las tasas es menor que al inicio del periodo. En específico, el spread entre las tasas de corto y largo plazo cayó 34 puntos básicos, mientras que el spread entre las tasas de mediano y de corto plazo apenas lo hizo en 22 puntos básicos.

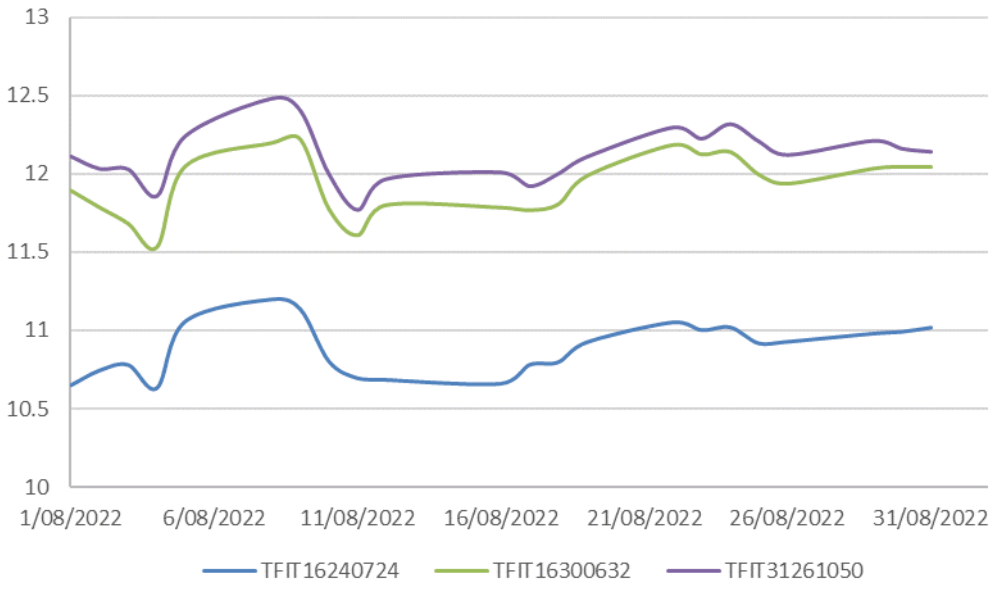
Las implicaciones en términos de riesgo de mercado de lo anterior van en el sentido de que la caída de las tasas en el largo plazo permitió una valorización o, en el peor de los casos, una menor pérdida, que lo que se registró para títulos de corto plazo. Es por esto por lo que, ante aplanamientos, lo mejor es comprar títulos de largo plazo, y vender títulos de corto plazo, siempre que en DV01 neto de la operación sea lo más cercano a cero posible. En caso de un empinamiento, entonces las decisiones deberían ser las contrarias, pero también cuidando el DV01 neto del portafolio.

Figura 7. Aplanamiento de la curva de rendimientos soberana colombiana entre fin de julio (rojo) y fin de agosto (azul) de 2022.



Tras la exploración de los datos, se concluye que todas las variables para todos los títulos tienen un nivel de importancia alto y no deben ser retiradas de los modelos. Finalmente, para entregarle a los modelos un set de datos uniforme, se incluye una columna adicional, en la que se entrega información de si el título se encuentra vigente o no, de modo que los modelos puedan descartar tomar en cuenta el título y sus datos faltantes cuando no está vigente, y solo tomando en cuenta los títulos que para determinada fecha se entran vigentes en el SEN.

Figura 8. Rendimientos TES corto, mediano y largo plazo durante aplanamiento de agosto 2022.



4.1.4 Modelación y Creación de Entornos de Aprendizaje por Refuerzo

El set de datos final es entregado a los modelos de aprendizaje por refuerzo A2C, DDPO y PPO, además de un modelo final de ensamble que busca seleccionar el modelo de mejor desempeño para una determinada ventana de tiempo. Este ensamble fue propuesto para el caso del mercado accionario estadounidense en Yang y otros (2020), mostrando resultados positivos superiores a los resultados obtenidos por los modelos trabajando individual e independientemente.

La ejecución de los modelos consta de ventanas de tiempo móviles, en las que el modelo de aprendizaje por refuerzo es entrenado con los datos más antiguos, dentro de un entorno de ejecución controlado de 63 días bursátiles, es decir, un trimestre calendario. Una vez entrenado el modelo, se ejecuta una fase de validación dentro de otro entorno de ejecución controlado también de 63 días bursátiles. Finalmente, un entorno de también 63 días es ejecutado con el fin de realizar la negociación de los títulos con datos no vistos antes; el en caso del modelo de ensamble, el modelo que será tomado en cuenta para esta etapa será aquel con mejor desempeño durante la etapa de validación, según la métrica de desempeño elegida. Una vez finalizada la iteración con esta ventana de ejecución, esta se mueve 63 días al futuro, y comienza nuevamente el ciclo de entrenamiento, validación y testeo.

Los tres entornos son muy similares en cuanto a su estructura. Cuentan con las siguientes

funciones principales:

- Inicialización: En esta función, se fija la fecha inicial y la partición del set de datos correspondiente a la ventana de tiempo con la que se va a trabajar. El espacio de observación se define con un tamaño total de 459 variables observables, mientras que el espacio de acción se limita al tamaño correspondiente a la cantidad de títulos que el agente puede comprar o vender, es decir, cincuenta. Se define el estado inicial y se inicializa el sistema de recompensas en cero.

  Se inicializa la ejecución en cualquier caso con los siguientes parámetros: Saldo inicial de la cuenta de diez mil millones de pesos colombianos. El coste de cada transacción efectuada por el agente corresponde al 0,1% del monto total de la misma, esto suponiendo que se tiene acceso como creador de mercado. Un factor de normalización de un millón de pesos, por lo que cada transacción deberá ser múltiplo de este número, como es normal en este mercado.

- Paso: Esta función toma los precios sucios de los títulos vigentes en determinado estado para ejecutar funciones de compra, venta o no realizar nada sobre estos. Finalizadas las transacciones, se actualiza el estado.

4.1.5 Métricas de Evaluación y Evaluación de los Modelos

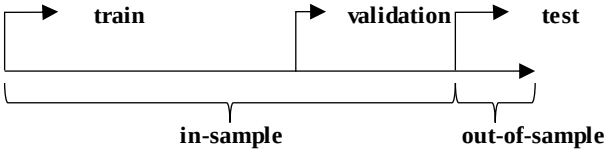
Con los tres entornos de ejecución creados, entonces es posible cargar los datos y ejecutar las secuencias de aprendizaje por refuerzo para los diferentes modelos para evaluarlos posteriormente. En ese sentido, las métricas de evaluación con las que serán comparados los modelos entre sí para elegir cuál de ellos será utilizado en el modelo de ensamble para una ventana de tiempo determinada, así como la evaluación final de los modelos, serán la utilidad neta total, es decir, en términos monetarios la cantidad de dinero que el modelo logró hacer para esa ventana de tiempo, y una versión más ligera de la razón de Sharpe (Sharpe, 1998), siendo esta la variable objetivo de este trabajo[6], pues su maximización permite mantener en un portafolio de inversión una relación sana entre la rentabilidad

[6] la razón de Sharpe ya ha sido propuesta como variable objetivo de maximización en problemas de Q-Learning por las ventajas que esta tiene para la generalización de la relación entre el retorno y el riesgo, generando resultados y rentabilidades satisfactorias, como las mostradas en Gao & Chan (2000).

porcentual obtenida sobre la variabilidad de los retornos diarios del portafolio.

Las ventanas de tiempo en las que los modelos trabajarán tendrán el siguiente esquema (También se puede observar en la figura 9): Se toman inicialmente dos trimestres[7] (luego se irán incrementando según la ventana de negociación se mueva) para entrenar los modelos A2C, DDPG y PPO, para que posteriormente estos modelos sean validados *in-sample* por los siguientes dos trimestres. Es en este punto en que se selecciona el modelo con mejor razón de sharpe en su etapa de validación, para hacer parte de la fase de negociación *out-of-sample* del modelo ensamblado. Adicionalmente también se prueban los modelos A2C, DDPG y PPO en la fase de negociación, la cual tiene una duración de un trimestre.

Figura 9. Esquema de partición de datos.



Finalmente, los modelos se evaluarán bajo los criterios de algunas otras métricas financieras comúnmente utilizadas en la gestión de portafolios, además de la razón de Sharpe, como pueden ser el retorno y la volatilidad total y anualizados, y máximo drawdown.

- Advantage Actor-Critic (A2C):

  Los resultados individuales de este modelo son bastante malos, en general. Con una razón de Sharpe total negativo, con un bajo promedio en esa misma métrica, y tan solo unos trimestres generando utilidades, la estrategia implementada por este tipo de modelo de actor-crítico generó resultados insatisfactorios, observables en la tabla 6, y fue superior a los demás modelos durante un solo trimestre, como se puede observar en la tabla 5. La utilidad total del modelo fue negativa, es decir, se materializaron pérdidas en el portafolio -cuyo balance inicial era de diez mil millones de pesos- por poco más

[7] Todos los trimestres son iguales en tamaño, es decir, no son trimestres calendario. En cambio, están compuestos por 62 días bursátiles consecutivos.

de seis millones de pesos.

- Proximal Policy Optimization (PPO):

  Los resultados individuales de este modelo son interesantes, pues se generaron utilidades por poco más de ciento treinta y cuatro millones de pesos, esto con una razón de Sharpe promedio de 0.70, esto es, que por cada unidad de riesgo de mercado asumida se generaron 0.70 unidades de retorno. Si bien en los primeros trimestres la razón de Sharpe fue modesta, en adelante los resultados fueron bastante importantes, lo que llevó a este modelo a ser elegido en quince de veintisiete trimestres como el mejor en el periodo de validación, para así hacer parte de la estrategia de ensamble. PPO generó un Sharpe positivo en el momento más álgido de la pandemia de 2020, sin embargo, se puede decir que el trimestre en el cual PPO operó con los datos de entrenamiento que venían de la pandemia, fue un año después de esta, y por ello, al igual que el modelo DDPG, no operó durante el primer trimestre de 2021.

- Deep Deterministical Policy Gradient (DDPG):

  Los resultados de este modelo fueron bastante modestos durante todo el periodo prepandemia, a excepción del segundo trimestre de negociación. Durante el periodo prepandemia, precisamente, este modelo pasó varios trimestres tomando la decisión de no operar como la mejor decisión para dicho trimestre, como se evidencia en la figura 10. Posteriormente, en la etapa pospandemia, DDPG fue el modelo que mejor incorporó la nueva dinámica de los mercados de renta fija, logran razones de Sharpe bastante relevantes, siendo la máxima de ellas de 17.8 veces, es decir, por cada unidad de riesgo de mercado asumido durante ese trimestre, se generaron 17.8 unidades de retorno. La razón de Sharpe total promedio fue de 2.12, y el modelo fue elegido para conformar el ensamble en once oportunidades, seis de ellas de manera consecutiva pospandemia. La utilidad total fue un poco superior a los ciento cuarenta y cuatro millones de pesos.

Tabla 5. Para cada periodo de 62 días bursátiles, el modelo elegido para la estrategia de ensamble es el de mayor razón de Sharpe.

Fecha Final Trim	Modelo Elegido	PPO	A2C	DDPG	Ensamble
20/04/2016	PPO	**0.93**	0.26	-0.28	0.93
22/07/2016	DDPG	0.50	0.06	**0.58**	0.58
20/10/2016	DDPG	-0.30	-0.39	**5.12**	5.12
23/01/2017	DDPG	-0.56	-0.42	**0.29**	0.29
24/04/2017	PPO	**0.29**	-0.24	0.00	0.29
27/07/2017	PPO	**0.77**	-0.32	-0.40	0.77
26/10/2017	PPO	**0.02**	-1.25	0.00	0.02
31/01/2018	DDPG	0.71	0.00	**1.26**	1.26
3/05/2018	PPO	**0.02**	-0.15	-0.06	0.02
6/08/2018	PPO	**2.13**	1.35	0.81	2.13
6/11/2018	A2C	0.21	**2.45**	0.00	2.45
7/02/2019	DDPG	1.21	0.20	**1.27**	1.27
10/05/2019	PPO	**0.19**	-0.10	0.00	0.19
12/08/2019	PPO	**0.02**	0.01	0.00	0.02
12/11/2019	PPO	**1.31**	0.04	0.00	1.31
13/02/2020	PPO	**1.02**	-0.05	0.00	1.02
15/05/2020	PPO	**1.70**	0.00	0.00	1.70
20/08/2020	PPO	**0.39**	-0.01	-0.03	0.39
19/11/2020	PPO	**0.43**	-0.05	0.00	0.43
23/02/2021	PPO	**0.00**	-0.12	0.00	0.00
26/05/2021	PPO	**1.25**	-0.05	0.00	1.25
27/08/2021	DDPG	1.21	0.36	**1.75**	1.75
26/11/2021	DDPG	2.08	0.94	**2.30**	2.30
25/02/2022	DDPG	0.52	1.94	**7.00**	7.00
27/05/2022	DDPG	0.41	1.75	**8.42**	8.42
31/08/2022	DDPG	1.03	0.28	**17.81**	17.81
30/11/2022	DDPG	1.39	0.49	**11.29**	11.29

Tabla 6. Resumen de métricas importantes de evaluación de modelos según su capacidad para administrar el portafolio entregado.

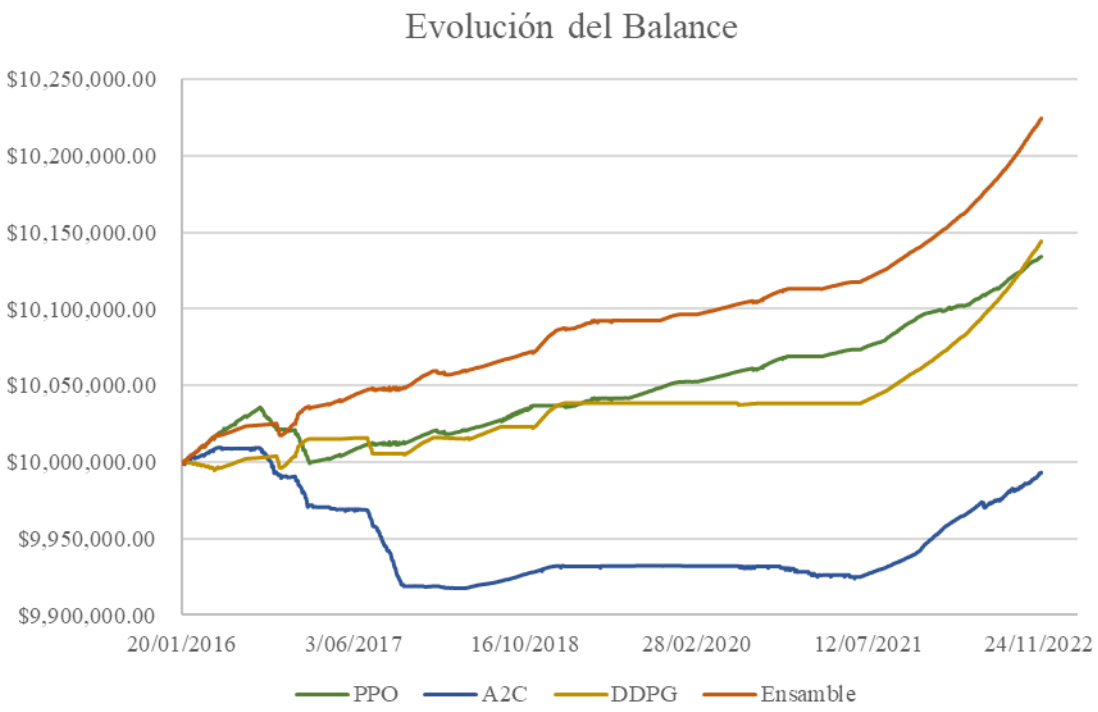
Métrica	PPO	A2C	DDPG	Ensamble
Sharpe Promedio	0.70	0.26	2.12	2.59
Sharpe Total	0.25	-0.01	0.40	0.52
Utilidad Total (Miles)	$ 134,629	-$ 6,656	$ 144,480	$ 225,044
Retorno Total	1.35%	-0.07%	1.44%	2.25%
Retorno Medio Anualizado	0.21%	-0.01%	0.22%	0.35%
Volatilidad Anualizada	0.05%	0.06%	0.03%	0.04%
Máximo Drawdown	-0.35%	-0.92%	-0.11%	-0.08%

Desde los resultados consolidados en la tabla 6, es posible identificar que el modelo A2C presenta mayor volatilidad que los otros dos modelos, y que su máximo drawdown es bastante superior al de los demás modelos, lo que explica su mal desempeño en términos de utilidad y retorno. En ninguna métrica, A2C fue superior a sus modelos pares, exceptuando un solo trimestre.

En el otro extremo, el modelo DDPG presenta niveles de volatilidad muy bajos, posiblemente provenientes de la gran cantidad de trimestres en los que el modelo no operó, que se evidencia en la tabla 5. Aun así, gracias a los excelentes resultados de este modelo en los últimos trimestres, la razón de Sharpe promedio de dicho modelo de 2.12 permite concluir que fue el mejor de los tres modelos. El hecho de que este modelo triplique en su razón de Sharpe a PPO, pero apenas le saque poco menos del 10% de ventaja en cuanto a utilidad total, tiene que ver con los efectos del interés compuesto dado que no se realizaron operaciones en los mencionados trimestres de inactividad.

PPO, en cambio, fue un modelo más moderado, con una razón de Sharpe promedio y total menor a uno, pero cauteloso en cuanto a volatilidad si se compara con A2C. Este modelo fue más constante que sus pares, pues solo en un trimestre el agente tomó la decisión de no operar, y, además, fue el único modelo en tener mejor desempeño que el modelo de ensamble por un corto periodo de tiempo, como se puede observar en la figura 10.

Figura 10. Comparación de la evolución del balance del portafolio según la estrategia implementada por los modelos PPO, A2C, DDPG y de ensamble.



Modelo de Ensamble

El objetivo de tener un modelo de ensamble es el de poder elegir específicamente cuál de los modelos anteriores utilizar para cierto periodo de tiempo, incorporando las características cambiantes del entorno de ejecución, en este caso, del mercado de bonos soberanos. Dichas características pueden ser la volatilidad, la liquidez o la tendencia que siguen los precios y las tasas de estos títulos. Para poder elegir, se requiere entonces una métrica que incorpore las características cambiantes parcial o totalmente, por lo que, en este caso, la razón de Sharpe es precisamente esa métrica que incorpora tendencia y volatilidad que puede cambiar con el tiempo en el mercado.

En este caso, según la razón de Sharpe de cada trimestre, es decir, con las condiciones de entrenamiento y validación previas al trimestre de la fase de negociación, se escoge el mejor de los tres modelos (A2C, PPO o DDPG) para hacer parte del modelo de ensamble para el trimestre de negociación, tal como se aprecia en la tabla 5.

Los resultados del modelo de ensamble fueron positivos en todas las métricas utilizadas. La rentabilidad fue la mejor de los cuatro modelos, la volatilidad asumida fue bastante baja, y el máximo drawdown fue el mejor, logrando así la mejor razón de Sharpe de los modelos evaluados.

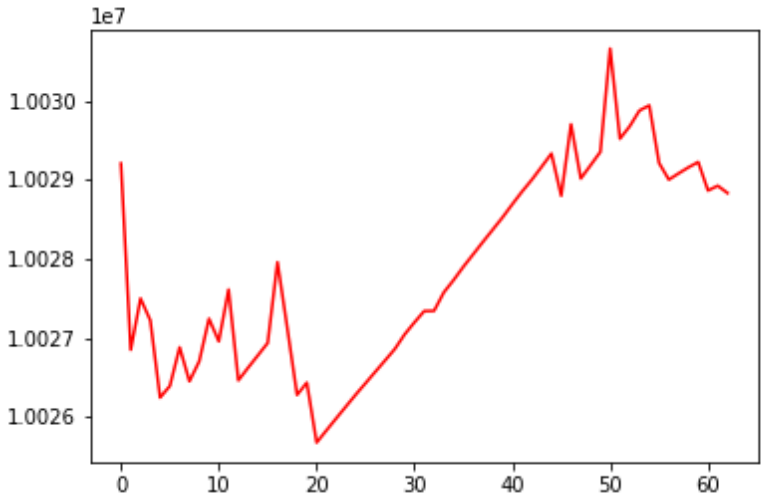
Es importante destacar que el modelo de ensamble, al integrar el mejor modelo en fase de validación por trimestre, en realidad está utilizando información más antigua, y, aun así, el criterio de selección del modelo basado en mayor razón de Sharpe, permite elegir un modelo que es capaz de generar rentabilidad para el siguiente trimestre, con información no antes vista y pese a que la situación del mercado pueda haber cambiado. De hecho, ante cambios en la coyuntura, el modelo de ensamble muestra capacidad de adaptación a esa nueva dinámica al escoger cuál de los modelos se adapta mejor, situación que se puede ejemplificar bastante bien al analizar conjuntamente la tabla 5 y la figura 9: Al cambiar la coyuntura económica con la pandemia de 2020, el modelo de ensamble eligió un cambio de modelo, y, por tanto, de esquema de operación -es decir, cambio de la función de optimización de la función de ventaja- ante las diferencias de los entornos de mercado prepandemia y pospandemia. Los demás modelos no pueden tener esta versatilidad al tener un único esquema de optimización.

Por último, en términos de rentabilidad, los resultados de todos los modelos, el de ensamble inclusive, si bien son positivos a excepción de A2C, en realidad no son suficientemente buenos. Esto pues, hablamos de una rentabilidad para siete años de muestra y seis de operación de 0,35% anual para el modelo de ensamble, el cual fue quien mostró mejor comportamiento en esta métrica.

La baja rentabilidad tiene varios orígenes: en primer lugar, a la instrucción de minimizar el riesgo vía razón de Sharpe, lo que, para el mercado de renta fija, significa buscar la minimización del DV01 del portafolio, de manera que las posiciones direccionales en puntos de la curva fueron contrarrestadas por posiciones contrarias en otros nodos, situación que es perfectamente normal en la gestión diaria de portafolios de renta fija. En segundo lugar, las 1,700 operaciones en promedio realizadas en los periodos de negociación no implican que el balance se usara completamente, por lo que la existencia de capital ocioso, e incluso de periodos sin operación, fueron bastante comunes, como se muestra en la figura

10, en donde para el décimo trimestre de operación, se observa una etapa de veinte días en los cuales el balance crecía uniformemente, indicando que en este lapso el modelo no tomó decisiones de inversión o desinversión. En la sección de discusiones y conclusiones se propone que pudiese hacerse con estos periodos de relativa inactividad.

Figura 11. Décimo trimestre de operación para el modelo de ensamble. Eje vertical es el balance del portafolio. Eje horizontal presenta el día de operación del trimestre.



Es importante señalar que, al igual que los demás modelos, el máximo drawdown del modelo de ensamble no se dio durante los fuertes movimientos en los mercados financieros provocados por la pandemia, ni siquiera se registró alguna caída para esas fechas. Esto es contrario a lo sucedido en (Yang y otros, 2020), en donde los autores se vieron obligados a utilizar una variable llamada «turbulencia» para evitar que sus estrategias operaran en días con condiciones de volatilidad extraordinarias. En cambio, en este trabajo, la baja volatilidad de los activos de renta fija -especialmente los calificados como AAA como el caso de los TES clase B-, sumado con la relación inversa entre tasa de negociación y precio limpio, y con el bajo uso de balance, permitió que, sin utilizar la «turbulencia», el modelo generase rentabilidad positiva durante esos días volátiles, evitando un drawdown, que de hecho solo se presentaron en los modelos durante los primeros meses de negociación, cuando los datos de entrenamiento eran muy pocos, como se evidencia en la figura 9.

A modo de anotación final sobre el modelo de ensamble, en el largo plazo siempre tuvo un balance superior al de cualquiera de los modelos individuales, mostrando su superioridad y aun así garantizando estabilidad y mejor precisión de largo plazo a la hora de tomar decisiones en el entorno de negociación sobre compras y ventas, «holdear» posiciones, e, incluso, sobre no usar el balance.

5. CONCLUSIONES Y DISCUSIÓN

En este trabajo se revisó el desempeño de diferentes modelos de aprendizaje por refuerzo, A2C, PPO y DDPG, y de un modelo ensamblado de estos en el contexto de la gestión activa de un portafolio de títulos de renta fija soberana colombiana, cumpliendo así con el objetivo principal planteado en la primera sección.

Los datos de mercado, provenientes del SEN del Banco de la República, fueron sometidos a un ciclo completo de ingeniería de datos, en donde se hizo una limpieza del set de datos, y una descripción completa de las variables que serían entregadas a los modelos, así como las transformaciones necesarias realizadas basadas en un buen entendimiento del negocio de administración de portafolios de renta fija y las necesidades puntuales que surgen en este tipo de mercados financieros.

Los modelos fueron sometidos a dos fases previas, una dentro de un ambiente de entrenamiento y la siguiente en un ambiente de validación, antes de llevarlos a ejecutar activamente su estrategia de administración, ajustando varios parámetros de mercado dentro de la ejecución del modelo para garantizar la mayor fidelidad a un entorno real de negociación, siendo estos: El tamaño del portafolio, unidades de normalización, costes transaccionales, y el tamaño de las ventanas temporales de entrenamiento y validación. La razón de Sharpe fue el indicador utilizado para buscar la optimización por parte de todos los modelos, es decir, para los modelos de aprendizaje reforzado, la razón de Sharpe era la recompensa o *reward* a maximizar, y fue escogido por su característica principal, esta es, la integración en un solo indicador entre la cantidad de utilidad generada por cada unidad de riesgo de mercado asumido, por lo que los modelos no se enfocarían en buscar una gran rentabilidad únicamente, sino también en mantener un portafolio de riesgo bajo a lo largo del tiempo.

Finalmente, los resultados de los modelos fueron evaluados desde una óptica financiera adecuada para las necesidades del negocio de administración de portafolios, con métricas e indicadores propios de esta industria. Se encuentra que, tras el entrenamiento, algunos modelos pueden elegir no operar durante el trimestre de validación, mostrando que son precavidos ante una situación específica del mercado que no lograron leer. En cambio, cuando los modelos si operan, pueden razones de Sharpe positivas asumiendo poco riesgo.

Sobre el modelo de ensamble, este escoge que modelo utilizar para un trimestre a partir de la razón de Sharpe que los modelos generan en la fase de validación. El modelo de ensamble escogió con frecuencia antes de pandemia operar bajo las instrucciones de PPO, y después de pandemia bajo las instrucciones de DDPG, lo que muestra que el modelo de ensamble, gracias a la flexibilidad que le brinda poder elegir el mejor de los modelos, tiene un mayor nivel de adaptación a los cambios que pueden darse en el mercado y que tal vez los modelos vistos de manera individual no pueden ver. Además, en el largo plazo, los resultados para el modelo de ensamble bajo las métricas e indicadores descritos fueron bastante superiores, con razones de Sharpe siempre positivas.

Si bien todos los objetivos del trabajo se cumplieron, es importante mencionar que hay muchos temas aun por abordar en posibles trabajos futuros, todos teniendo en cuenta que, si bien se logró generar una razón de Sharpe interesante, esto fue por la baja cantidad de riesgo de mercado asumido en todas las estrategias, con, de hecho, un muy bajo uso de balance.

En primer lugar, puede ser objetivo de un trabajo futuro buscar motivar a los modelos a asumir algo más de riesgo de mercado o usar mejor su balance. Esto puede inducirse al entregarle a los modelos más acciones posibles a realizar, pues para este trabajo, los modelos solo tenían disponible las acciones de comprar, vender o mantener los títulos. En el mercado de renta fija, y especialmente el soberano dada su alta liquidez, es posible realizar otro tipo de operaciones llamadas operaciones de liquidez o simultáneas. Agregar este tipo de operaciones al modelo para mejorar el uso de balance implica hacer todo el ciclo de ingeniería de datos para las operaciones que se realizan diariamente en la rueda de simultáneas del SEN.

Por otro lado, también existe interés en que los modelos incorporen correctamente estrategias completas sobre la curva de rendimientos. Si bien los modelos en este trabajo lograron rentabilidad operando diferentes instrumentos en diferentes nodos de la curva, se puede buscar la manera de potenciar estos resultados, tanto en términos de rentabilidad como de riesgo, buscando modelos que realicen predicciones precisas sobre los movimientos de la curva de rendimientos o sobre los parámetros de la curva de Nelson y Siegel (1987).

En último lugar, y pensando el despliegue de los modelos para que los administradores de portafolios puedan utilizarlo, se debe evaluar el esquema de costos que puede tener dicho administrador para poder operar en ese mercado, y que puede variar con respecto al costo de

operación utilizado para este trabajo, pues es un costo por operación para una entidad que se encuentre dentro del esquema de creadores de mercado. Adicionalmente, hay un costo de transacción que no influye en la precisión de los modelos, pero si en su futura rentabilidad, y este es el bid-ask spread, es decir, la diferencia entre los precios de compra y de venta de un mismo título. Este costo no puede ser tenido en cuenta en los modelos por su alta dependencia a las condiciones del mercado, pero para el despliegue de los modelos en, por ejemplo, un aplicativo para los administradores, este costo adicional puede llegar a disminuir la rentabilidad que los modelos esperan generar.

Este trabajo fue una buena aproximación a un problema que no ha sido ampliamente abordado, pues los autores suelen enfocar sus esfuerzos en activos de renta variable. Los trabajos futuros que resulten de este trabajo inicial deben propender por la sofisticación de los modelos para mejorar la razón de Sharpe u otros indicadores de relevancia, y por incorporar más información del mercado, como pueden ser las predicciones a los movimientos de la curva de rendimientos.

Agradecimientos

Agradecimientos especiales a mi directora de trabajo de grado, Paula Almonacid, por su instrucción, disposición y acompañamiento durante todo el proceso de investigación y construcción del presente trabajo. A la Universidad EAFIT, al ICETEX y al Ministerio de las TIC, por su apoyo académico y financiero durante todo el posgrado. A mi familia por su incondicional apoyo y su inconmensurable cariño.

Finalmente, establecer una dedicatoria de este trabajo de grado a mi pareja, Stefania Restrepo, pues pienso que de su parte recibí toda la motivación necesaria, y más, para finalizar este trabajo.

Bibliografía

Ang, S., Alles, L., & Allen, D. (1998). Riding the yield curve: An analysis of international evidence. *The Journal of Fixed Income, 8*(3), 57-74.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine, 34*(6), 26-38. https://doi.org/10.1109/MSP.2017.2743240

Banco de la República. (12 de 2022). *Estadísticas SEN: Banco de la República.* Banco de la República: https://www.banrep.gov.co/es/sistemas-pago/estadisticas-sen

Bellman, R. (1952). On the Theory of Dynamic Programming. *Proceedings of the National Academy of Sciences, 38*(8), 716-719. https://doi.org/10.1073/pnas.38.8.716

Bellman, R. (1966). Dynamic programming. *Science, 153*(3731), 34-37. https://doi.org/10.1126/science.153.3731.34

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of political economy, 81*(3), 637-654.

Bolsa de Valores de Colombia S.A. (2023). *Mercado Local en Línea*. Bolsa de Valores de Colombia: www.bvc.com.co

Busoniu, L., Babuska, R., & De Schutter, B. (2008). A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38*(2), 156-172. https://doi.org/10.1109/TSMCC.2007.913919

Carapuco, J., Ferreira-Neves, R., & Horta, N. C. (2018). Reinforcement learning applied to forex trading. *Applied Soft Computing, 73*, 783-794. https://doi.org/10.1016/j.asoc.2018.09.017

Carta, S., Rerreira, A., Podda, A. S., Recupero, D. R., & Sannai, A. (2021). Multi-dqn: An ensemble of deep q-learning agents for stock market forecasting. *Expert Systems with Applications, 164*, 113820. https://doi.org/10.1016/j.eswa.2020.113820

Chua, C. T., Koh, W. T., & Ramaswamy, K. (2005). Comparing returns of US treasuries versus equities: implications for market and portfolio efficiency. *Applied Financial Economics, 15*(17), 1213-1218.

Dixon, M., Halperin, I., & Bilokon, P. (2020). *Machine Learning in Finance: From Theory to Practice.* Springer International.

Dyl, E., & Joehnk, M. D. (1981). Riding the yield curve: does it work? *The journal of portfolio management, 7*(3), 13-17.

Fabozzi, F. J. (2007). *Fixed income analysis.* John Wiley & Sons.

Fabozzi, F. J. (2021). *The handbook of fixed income securities* (Novena ed.). McGraw-Hill Education.

Fernández-Rodríguez, F., González-Martel, C., & Sosvilla-Rivero, S. (2000). On the profitability of technical trading rules based on artificial neural networks: Evidence from the Madrid stock market. *Economics Letters, 69*(1), 89-94. https://doi.org/https://doi.org/10.1016/S0165-1765(00)00270-6

Galvani, V., & Landon, S. (2013). Riding the yield curve: a spanning analysis. *Review of Quantitative Finance and Accounting, 40*, 135-154.

Gao, X., & Chan, L. (2000). An algorithm for trading and portfolio management using q-learning and sharpe ratio maximization. *Proceedings of the international conference on neural information processing* (págs. 832-837). Hong Kong: Citeseer.

Gogas, P., Papadimitriou, T., Matthaiou, M., & Chrysanthidou, E. (2015). Yield curve and recession forecasting in a machine learning framework. *Computational Economics, 45*, 635-645.

Grieves, R., & Marcus, A. J. (1992). Riding the yield curve: reprise. *Journal of Portfolio Management, 18*(4), 67-76.

Gu, S., Lillicrap, T., Sutskever, I., & Levine, S. (2016). Continuous Deep Q-Learning with Model-based Acceleration. *Proc. Int. Conf. Learning Representations,.* https://arxiv.org/abs/1603.00748

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications, 124*, 226-251. https://doi.org/https://doi.org/10.1016/j.eswa.2019.01.012

Hull, J. C. (2022). *Options, Futures, and Other Derivatives* (Undécima ed.). Pearson.

Kakade, S. M. (2001). A Natural Policy Gradient. *Advances in Neural Information Processing Systems* (págs. 1531-1538). Cambridge: MIT Press.

Konda, V. R., & Tsitsiklis, J. N. (2003). OnActor-Critic Algorithms. *SIAM Journal on Control and Optimization, 42*(4), 1143-1166. https://doi.org/10.1137/S0363012901385691

Konda, V., & Tsitsiklis, J. (1999). Actor-Critic Algorithms. *Advances in Neural Information Processing Systems.* Boston: MIT Press. https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Koutník, J., Cuccu, G., Schmidhuber, J., & Gomez, F. J. (2013). Evolving large-scale neural networks for vision-based reinforcement learning. *Annual Conference on Genetic and Evolutionary Computation.* Manno. https://doi.org/10.1145/2463372.2463509

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436-444. https://doi.org/10.1038/nature14539

Li, S., Wu, Y., Cui, X., Dong, H., Fang, F., & Russell, S. (2019). Robust Multi-Agent Reinforcement Learning via Minimax Deep Deterministic Policy Gradient. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 4213-4220. https://ojs.aaai.org/index.php/AAAI/article/view/4327

López de Prado, M. M. (2020). *Machine Learning for Asset Managers.* Cambridge University Press. https://doi.org/10.1017/9781108883658

Merton, R. C. (1973). Theory of Rational Option Pricing. *The Bell Journal of Economics and Management Science, 4*(1), 141-183.

Moody, J., & Saffell, M. (1998). Reinforcement Learning for Trading. En M. Kearns, S. Solla, & D. Cohn (Ed.), *Advances in Neural Information Processing Systems. 11*. Boston: MIT Press. https://proceedings.neurips.cc/paper_files/paper/1998/file/4e6cd95227cb0c280e99a195be5f6615-Paper.pdf

Nelson, C. R., & Siegel, A. F. (1987). Parsimonious Modeling of Yield Curves. *Journal of Business, 60*(4), 473-489.

Nunes, M. (2022). *Machine learning in fixed income markets: forecasting and portfolio management.* University of Southampton.

Pelaez, R. F. (1997). Riding the yield curve: Term premiums and excess returns. *Review of Financial Economics, 6*(1), 113-119. https://doi.org/https://doi.org/10.1016/S1058-3300(97)90017-3

Refenes, N. A.-P., Burgess, N. A., & Bentz, Y. (1997). Neural networks in financial engineering: a study in methodology. *IEEE transactions on neural networks, 8*, 1222-1267. https://doi.org/10.1109/72.641449

Schulman, J., Levine, S., Moritz, P., Jordan, M., & Abbeel, P. (2015). Trust Region Policy Optimization. *International conference on machine learning*, 1889-1897. https://doi.org/https://doi.org/10.48550/arXiv.1502.05477

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. *OpenAI*. https://doi.org/arXiv:1707.06347

Sharpe, W. (1998). The Sharpe Ratio. *Streetwise, 3*, 169-185.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014). Deterministic Policy Gradient Algorithms. *Proceedings of Machine Learning Research.* Beijing: PMLR. https://proceedings.mlr.press/v32/silver14.html

Stone, P., & Kohl, N. (2004). Policy gradient reinforcement learning for fast quadrupedal locomotion. *IEEE International Conference on Robotics and Automation, 3*, 2619-2624. https://doi.org/10.1109/ROBOT.2004.1307456

Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning.* University of Massachusetts Amherst.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (Segunda ed.). MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence, 112*(1), 181-211. https://doi.org/https://doi.org/10.1016/S0004-3702(99)00052-1

Tao, C., & Wencong, S. (2018). Local Energy Trading Behavior Modeling With Deep Reinforcement Learning. *IEEE Access, 6*, 62806-62814. https://doi.org/10.1109/ACCESS.2018.2876652

van Hasselt, H., Guez, A., & Silver, D. (2015). Deep Reinforcement Learning with Double Q-Learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 30.* https://doi.org/10.1609/aaai.v30i1.10295

Wang, X., Wang, S., Liang, X., Zhao, D., Huang, J., Xu, X., . . . Miao, Q. (2022). Deep reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1-15. https://doi.org/10.1109/TNNLS.2022.3207346

Watkins, C. J. (1989). *Learning from delayed rewards.* Learning from delayed rewards.

Watkins, C. J., & Dayan, P. (1992). Q-Learning. *Machine Learning, 8*(3), 279-292. https://doi.org/10.1007/BF00992698

Wu, X., Chen, H., Wang, J., Troiano, L., Loia, V., & Fujita, H. (2020). Adaptive stock trading strategies with deep reinforcement learning methods. *Information Sciences, 538*, 142-158.

Yang, H., Lui, X.-Y., Zhong, S., & Walid, A. (2020). Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy. *Proceedings of the First ACM International Conference on AI in Finance.* Nueva York: Association for Computing Machinery. https://doi.org/10.1145/3383455.3422540

Yoon, Y., Swales, G., & Margavio, T. M. (1993). A Comparison of Discriminant Analysis versus Artificial Neural Networks. *Journal of the Operational Research Society, 44*(1), 51-60. https://doi.org/10.1057/jors.1993.6

Zenios, S. A., & Ziemba, W. T. (2007). *Handbook of Asset and Liability Management: Applications and case studies.* Elsevier.

Zhang, N., Lin, A., & Shang, P. (2017). Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting. *Physica A: Statistical Mechanics and its Applications, 477*, 161-173. https://doi.org/https://doi.org/10.1016/j.physa.2017.02.072

Zimmermann, H., Neuneier, R., & Grothmann, R. (2000). Modeling of the german yield curve by error correction neural networks. *Proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning, Data Mining, Financial Engineering, and Intelligent Agents* (págs. 262-267). Berlín: Springer-Verlag.