INSTITUTO TECNOLÓGICO DE CIUDAD MADERO

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

TESIS

APRENDIZAJE POR REFUERZO PROFUNDO PARA TRADING ALGORITMICO

Que para obtener el Grado de:
Maestro en Ciencias de la Computación

Presenta
I.S.C. Irving Omar Rodríguez Hernández
G14071343
No. de CVU 1037457

Director de Tesis:
Dr. Juan Javier González Barbosa
No. de CVU 202134

Co-director de Tesis:
Dr. Juan Frausto Solís

Ciudad Madero, Tamaulipas. mayo 2022

Instituto Tecnológico de Ciudad Madero
Subdirección Académica
División de Estudios de Posgrado e Investigación

Ciudad Madero, Tamaulipas, 16/marzo/2022

OFICIO No.: U.041/22
ASUNTO: AUTORIZACIÓN DE IMPRESIÓN DE TESIS

C. IRVING OMAR RODRÍGUEZ HERNÁNDEZ
No. DE CONTROL G14071343
P R E S E N T E

Me es grato comunicarle que después de la revisión realizada por el Jurado designado para su Examen de Grado de Maestría en Ciencias de la Computación, se acordó autorizar la impresión de su tesis titulada:

"APRENDIZAJE POR REFUERZO PROFUNDO PARA TRADING ALGORITMICO"

El Jurado está integrado por los siguientes catedráticos:

PRESIDENTE:	DRA.	GUADALUPE CASTILLA VALDEZ
SECRETARIO:	DRA.	LAURA CRUZ REYES
VOCAL:	DR.	JUAN JAVIER GONZÁLEZ BARBOSA
SUPLENTE:	DR.	JUAN FRAUSTO SOLÍS
DIRECTOR DE TESIS:	DR.	JUAN JAVIER GONZÁLEZ BARBOSA
CO-DIRECTORA:	DR.	JUAN FRAUSTO SOLÍS

Es muy satisfactorio para la División de Estudios de Posgrado e Investigación compartir con usted el logro de esta meta. Espero que continúe con éxito su desarrollo profesional y dedique su experiencia e inteligencia en beneficio de México.

A T E N T A M E N T E
Excelencia en Educación Tecnológica.
"Por mi patria y por mi bien".

MARCO ANTONIO CORONEL GARCÍA
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

INSTITUTO TECNOLÓGICO DE CIUDAD MADERO
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

ccp. Archivo
MACG

Av. 1° de Mayo y Sor Juana I. de la Cruz S/N Col. Los Mangos C.P. 89440 Cd. Madero, Tam.
Tel. 01 (833) 357 48 20, ext. 3110, e-mail: depi_cdmadero@tecnm.mx
tecnm.mx | cdmadero.tecnm.mx

Declaración de originalidad

Declaro y prometo que este documento de tesis es producto de mi trabajo original y que no infringe los derechos de terceros, tales como derechos de publicación, derechos de autor, patentes y similares. Por lo tanto, la obra es de mi autoría y soy titular de los derechos que surgen de la misma.

Declaro también que en las citas textuales que he incluido (las cuales aparecen entre comillas) y en los resúmenes que he realizado de publicaciones ajenas, indico explícitamente los datos de los autores y publicaciones.

Además, en caso de presentarse cualquier reclamación o acción por parte de un tercero en cuanto a los derechos de autor sobre la obra en cuestión, acepto toda la responsabilidad de tal infracción y relevo de ésta a mi director y codirector de tesis, así como al Tecnológico Nacional de México, al Instituto Tecnológico de Ciudad Madero y a sus respectivas autoridades.

Irving Omar Rodríguez Hernández

AGRADECIMIENTOS

A mi director de tesis, Dr. Juan Javier González Barbosa, por su paciencia y guía para el desarrollo y culminación de esta tesis.

A mi codirector, Dr. Juan Frausto Solis, por el tiempo invertido y su disposición para compartir el vasto conocimiento que posee en múltiples áreas de la ciencia e ingeniería.

A los miembros de mi comité, Dra. Laura Cruz Reyes, Dra. Guadalupe Castilla Valdez y Dr. Luis Fortino Cisneros Sinencio, por sus valiosos comentarios y sugerencias.

A mi madre, Ma. del Rocio Hernandez Cabañas, por su amor y apoyo incondicional.

A Selene, por su comprensión y compañía en los momentos difíciles.

Por último, agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca otorgada para realizar mis estudios de maestría y al Instituto Tecnológico de Cd. Madero por las facilidades proporcionadas para el desarrollo de este proyecto.

Aprendizaje por refuerzo profundo para Trading Algorítmico

Ing. Irving Omar Rodriguez Hernandez

Resumen

El problema de asignación continua de recursos en mercados financieros o trading conlleva la construcción de portafolios de inversión a ejecutar en intervalos de tiempo discretos. La solución a este problema exige el desarrollo de estrategias de comercialización de bienes que tengan como objetivos el incremento sustancial del capital inicial y una baja volatilidad (riesgo) en comparación con métodos y técnicas ya conocidos tanto para la gestíón activa como pasiva de portafolios de inversión.

Este trabajo presenta el método GA-DRL para la generación de estrategias de inversión con acciones de los mercados de valores. Por una parte, integra el algoritmo GenPo-Sharpe para la preselección de activos financieros cuya finalidad es proporcionar un portafolio de inversión con un número reducido de activos, de acuerdo a las preferencias del inversionista, y balanceado en cuanto a su rendimiento y riesgo esperados. En segundo lugar integra un agente de aprendizaje por refuerzo profundo para el rebalanceo diario del portafolio de inversiones obtenido en la primera etapa.

La concatenación de estas dos técnicas, que constituyen al método presentado, se provee como una alternativa adicional para la gestión activa de portafolios de inversión con resultados prometedores en acciones de las bolsas de valores de Nueva York y Brasil.

Deep reinforcement learning for Algorithmic Trading

Eng. Irving Omar Rodriguez Hernandez

Abstract

The continuous resources allocation on financial markets problem or trading implies building investment portfolios to be executed on discret interval times. Solving this problem requires the develoment of strategies for market operation whose aim is to increase the initial capital notably and reduce the volatilty (risk) in comparison with wide known techniques and methods for active and passive investment portafolio management.

This work presents GA-DRL method for investment strategies generation by using stock market financial instruments. On the one hand, this method integrates the GenPo-Sharpe algorithm for stock preselection having a number-constrained and expected performance risk balanced ratio investment portafolio as result, on the other a deep reinforcement learning agent is integrated for daily trading of the portfolio obtainted in the first stage.

The joint of this couple of techniques, that constitutes the GA-DRL method, is provided as an extra alternative for active investment portfolio management exposing promising results on New York and Brazil stock exchanges.

Tabla de Contenido

Índice de Tablas

Índice de figuras

1 Introducción

Los mercados financieros son uno de los tres componentes estructurales del sistema financiero cuyo rol es clave en la economía debido a que estimula su crecimiento y bienestar. De acuerdo al enfoque funcional, los mercados financieros facilitan el flujo de fondos mediante el financiamiento de inversiones por corporaciones, gobiernos e individuos (Darškuvienė, 2010)

En los mercados financieros es posible el *trading* (comercialización) de tres tipos de activos diferentes; materias primas, monedas (mercados FX) y valores.

Con el objetivo de que un tomador de decisiones (DM) financiero maximice las utilidades de sus inversiones este debe diseñar una estrategia de trading que, de acuerdo a diversas variables, como lo puede ser el historial de precios, distribuya de manera apropiada el capital disponible a los activos con mayor de probabilidad de revalorización a futuro. La complejidad de las actividades a realizar por el DM está determinada principalmente por el gran número de portafolios de inversión factibles a evaluar; una planeación o diseño de estrategias de inversión erradas pueden provocar la perdida total del capital.

La tecnología juega desde hace varias décadas un papel relevante en el área de finanzas. En el año 2018, más del 70% de las ordenes del mercado de derivados *Chicago Mercantile Exchange* (CME) se realizaron por robots (Meyer, 2019). Una sub área de finanzas, donde la tecnología esta fuertemente implicada, es el trading algorítmico (AT). AT se refiere a la aplicación de algoritmos computarizados en la automatización de una parte o todo el ciclo de trading. En consecuencia, AT involucra la aplicación de técnicas para el aprendizaje, la planeación dinámica, el razonamiento y la toma de decisiones (Treleaven, Galas, & Lalchand, 2013).

Este trabajo se centra en tres, de un total de cinco etapas que componen a los sistemas de AT (Treleaven, Galas, & Lalchand, 2013), investigación y recopilación de datos, análisis pre-trading y generación de señales de trading. De manera general, la primera etapa consiste en la determinación de las variables del mercado que alimentaran al sistema, la segunda etapa en la creación de modelos de ganancias y riesgos, por último, la tercera etapa involucra la

construcción de portafolios de inversión para los intervalos de tiempo especificados por un DM.

En el campo de *machine learning* (ML), tradicionalmente, la asignación continua de recursos financieros ha sido abordada mediante técnicas de pronostico de precios junto a reglas de trading definidas por las preferencias de un DM. En años recientes y debido a los notables resultados que ha tenido el paradigma de aprendizaje por refuerzo (RL) en la generación de estrategias de acción para sistemas complejos como videojuegos, juegos de mesa, etc. los investigadores han empleado, con resultados prometedores, agentes de RL en sistemas de AT.

Un agente artificial de RL adquiere conocimiento mediante la interacción con su entorno, teniendo como fin desarrollar una estrategia de comportamiento óptima. Los aspectos que distinguen a RL de otros paradigmas de ML son: la recompensa/castigo que se propaga en el tiempo y el mecanismo de acción de prueba-error que permite a un agente explorar y explotar su entorno. (Sutton & Barto, 2018).

Uno de los problemas a lidiar en la aplicación de RL en sistemas complejos, como lo son los mercados financieros, es el manejo del espacio de estados-decisiones multidimensional e incluso con componentes continuos. La técnica comúnmente empleada, como solución a este problema, son las redes neuronales profundas (DNN) en el papel de aproximadores de funciones; realizando la transformación de un agente de RL convencional a un agente de RL profundo (DRL).

El presente trabajo propone el método GA-DRL para trading de activos de la bolsa de Nueva York (NYSE). Este método consiste en un agente de DRL que es alimentado de múltiples series de tiempo derivadas de activos seleccionados mediante el algoritmo genético GenPo-Sharpe y se desenvuelve en dos etapas; una primera etapa de aprendizaje, utilizando el histórico de precios, volúmenes de comercialización e indicadores técnicos, que da como resultado estrategias de inversión y una segunda etapa para la evaluación de dichas estrategias. Como trabajo complementario, se evalúa el comportamiento del agente de DRL propuesto con activos de la bolsa de valores de Brasil (BSE).

1.1 Planteamiento del problema

El problema de asignación continua de recursos o trading en mercados financieros implica el diseño de estrategias de inversión que evolucionan acorde a los datos recopilados en cada intervalo de tiempo, con el objetivo de que un inversionista obtenga el mayor beneficio posible a partir de una serie de toma de decisiones de trading.

Teniendo en cuenta la naturaleza de toma de decisiones secuenciales del problema y la maximización del retorno a largo plazo esperada por el DM, se propone la formulación del problema de asignación continua de recursos financieros como un proceso de decisión Markov (MDP) para el que se busca obtener la política de trading óptima. Los componentes de un MDP se definen como una tupla $M = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$ donde, para este caso, las características del mercado (precios e indicadores técnicos) representan los estados del sistema $\mathcal{S}$, los portafolios factibles a ejecutar son el conjunto de decisiones disponibles $\mathcal{A}$ y la señal de recompensa $\mathcal{R}$ es una función para evaluar el rendimiento de las decisiones tomadas por el agente para cierto momento en el tiempo.

$$\max_{\pi \in \Pi} \mathbb{E}\left[\sum_{k=0}^{T} \gamma^k R\big(s_k, \pi(s_k)\big) \mid s_0\right] \tag{1}$$

Donde:

π: política perteneciente al conjunto total de políticas Π.
k: índice de un momento en el tiempo.
s_k: estado en el momento k.
$\pi(s_k)$: decisión a tomar en el estado s_k siguiendo la política π.
R: función de recompensa dado el estado s_k y la política π.
s_0: estado inicial.

Los MDPs pueden ser resueltos de manera óptima con técnicas de programación dinámica, sin embargo, dado el carácter multidimensional tanto de el espacio de decisiones como el espacio de estados y el desconocimiento de la dinámica del sistema; se plantea la solución mediante un agente de DRL. Entre los retos mas notables se encontraron; la definición de los componentes de los estados del sistema, ya que estos representan la información disponible por el agente

para la toma de decisiones y la arquitectura de la red neuronal profunda (DNN) que determinará las decisiones a tomar por nuestro agente.

1.2 Justificación

El problema de asignación continua de recursos o trading en mercados financieros ha sido tradicionalmente abordado desde tres enfoques; el análisis de series de tiempo financieras utilizando modelos basados en *autoregressive moving* average (ARMA), técnicas de ML como Redes neuronales artificiales (ANN) o Máquinas de soporte vectorial (SVM) y recientemente, debido al desarrollo de aprendizaje profundo (DL), por medio de redes neuronales recurrentes (RNN).

Estos enfoques presentan los siguientes inconvenientes:

- La generalización suele ser una técnica poco adecuada para la tarea de toma de decisiones secuenciales en entornos no-estacionarios.
- La generación de señales de trading no se realiza en base al comportamiento de los datos ni la experiencia del agente en el proceso, si no que son producto de heurísticas aplicadas por un DM.

En consecuencia y derivado de las técnicas antes mencionadas, un agente inversionista toma decisiones potencialmente ineficaces que generan ganancias a largo plazo por debajo de los índices de valores e incluso afrontando perdida de capital.

En virtud de resolver las deficiencias antes mencionadas; se identifica el paradigma de RL con DL, cuyas técnicas tienen el objetivo de aprender la dinámica de entornos complejos en base a la toma de decisiones secuenciales y de diseñar estrategias de decisión (políticas) para el cumplimiento de una meta en base a una señal de recompensa numérica.

Establecidas las desventajas de los métodos de trading algorítmico tradicionalmente utilizados y las bondades de RL y DL para el diseño de políticas de acción en sistemas complejos como los mercados financieros, este trabajo de investigación propone desarrollar un agente de DRL

para el diseño de estrategias de inversión que generen ganancias netas por encima de estrategias de trading habituales y ampliamente utilizadas como *Buy & Hold*.

1.3 Objetivos del proyecto

1.3.1 Objetivo general

Desarrollar un agente de aprendizaje por refuerzo profundo para trading algorítmico que genere ganancias superiores a la estrategia *Buy & Hold* en instrumentos pertenecientes a un mercado financiero regulado.

1.3.2 Objetivos específicos

Con motivo de cumplir el objetivo general, se presentan los siguientes objetivos específicos.

- Identificación del mercado de valores donde se desarrollará el agente.
- Adaptación del algoritmo GenPo-Sharpe para la preselección de instrumentos financieros.
- Obtención y preparación de las series de tiempo financieras.
- Establecimiento de los componentes del MDP: estados, decisiones y función de recompensa.
- Establecimiento del algoritmo de RL a utilizar.
- Establecimiento de la arquitectura de la DNN a incorporar en el agente de RL.
- Experimentación computacional del agente desarrollado.
- Exposición y comparación de resultados.

1.4 Alcances y limitaciones

A continuación, se enlistan los alcances:

- Emplear librerías o *frameworks* existentes para DRL; PyTorch, Keras, SLM Lab, etc.
- Realizar trading algorítmico para portafolios de máximo diez instrumentos financieros.
- Experimentar con conjuntos de activos pertenecientes a alguno de los siguientes sectores financieros: energía, tecnología o salud.
- Probar el agente de DRL en periodos no utilizados para la preselección de los activos ni el entrenamiento del agente.

Entre las limitaciones se contemplan:

- El uso de series de tiempo financieras (precios históricos o indicadores técnicos) como datos de entrada.
- Definición de la longitud de las series de tiempo no mayor a cinco años.
- Obtención de los instrumentos financieros que compondrán el portafolio de inversiones mediante el algoritmo GenPo-Sharpe.
- Series de tiempo financieras de un registro por día.

1.5 Organización de la tesis

El desarrollo de esta tesis se realiza en cinco capítulos. En el capitulo I, se menciona tanto la introducción y planteamiento del problema como los objetivos del proyecto. En el capitulo II se abordan los conceptos del marco teórico necesarios para comprender tanto la naturaleza del problema de asignación continua de recursos financieros como el método de DRL propuesto. A lo largo del capitulo III se hace una breve descripción de trabajos relacionados a esta tesis, En el capitulo IV se expone como un diagrama el método propuesto y el pseudocódigo del agente inversionista de DRL. En el capítulo V se muestra la evaluación del agente inversionista derivada de una experimentación computacional. Por último, en el capitulo VI se presentan las conclusiones y propuestas para trabajos de investigación futuros.

2 Marco Teórico

2.1 Mercados financieros

Los mercados financieros describen cualquier agregado de posibles compradores y vendedores de activos junto a las transacciones que existen entre ellos. La transparencia de los precios, las reglas de transacción, los costos y comisiones son características de los mercados regulados (Branco, 2017). Debido a que los activos comercializados en los mercados financieros son intangibles, el valor de los mismos es derivado a partir de contratos. Los activos comercializados pueden ser clasificados en tres grupos principales: materias primas, valores y monedas.

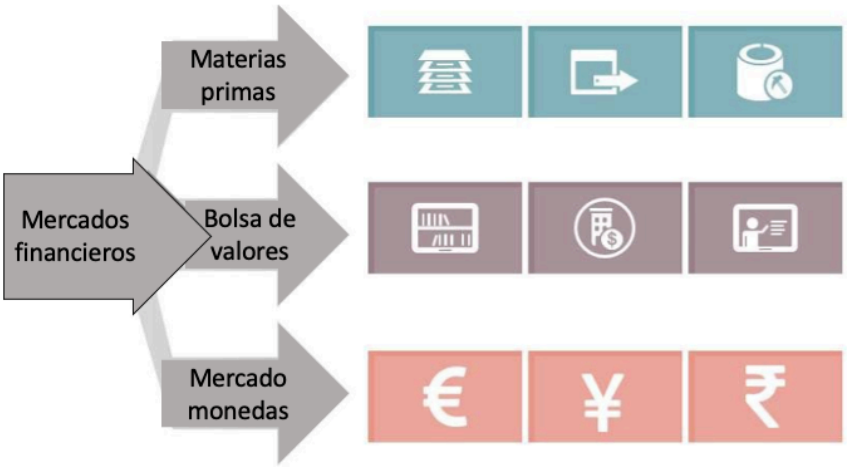

Figura 1 Tipos de mercados financieros

Un mercado financiero del primer grupo involucra la promesa de entrega de algún bien como lo pueden ser el oro, la plata, el petróleo, u otras materias primas.

En los mercados de valores encontramos un par de subcategorías: valores de deuda y títulos. Los valores de deuda representan dinero que ha sido puesto en préstamo con algunas condiciones de devolución (por ejemplo; una tasa de interés). Los títulos avalan la propiedad parcial de alguna entidad; en esta subcategoría se consideran los futuros, derivados y las acciones; uno de los activos comúnmente relacionados con el trading financiero y empleados en esta tesis para la experimentación computacional del agente de DRL propuesto

Por último, en el mercado de monedas se comercializan representaciones de dinero soportadas por los bancos centrales (Peso mexicano, Dólar americano, Euro, etc.). Una de las propiedades ha cumplir por las monedas comercializadas es la convertibilidad; es decir, el cambio a alguna otra moneda disponible en el mercado donde el inversionista se encuentre operando.

Con el objetivo de generar utilidades en los mercados financieros es imprescindible que un DM emplee herramientas y métodos de análisis que permitan actuar acorde al comportamiento presente y futuro del mercado. Los dos enfoques principales de análisis son: el análisis fundamental, basado en indicadores que describen el valor subyacente del bien y el análisis técnico cuyos indicadores se sustentan en la premisa de que el rendimiento histórico del mercado permite predecir su rendimiento futuro.

2.1.1 Trading financiero

La comercialización de activos dentro de un mercado financiero (trading financiero), puede ser ejecutado de acuerdo a cuatro enfoques principales: cobertura, arbitraje, inversión o especulación.

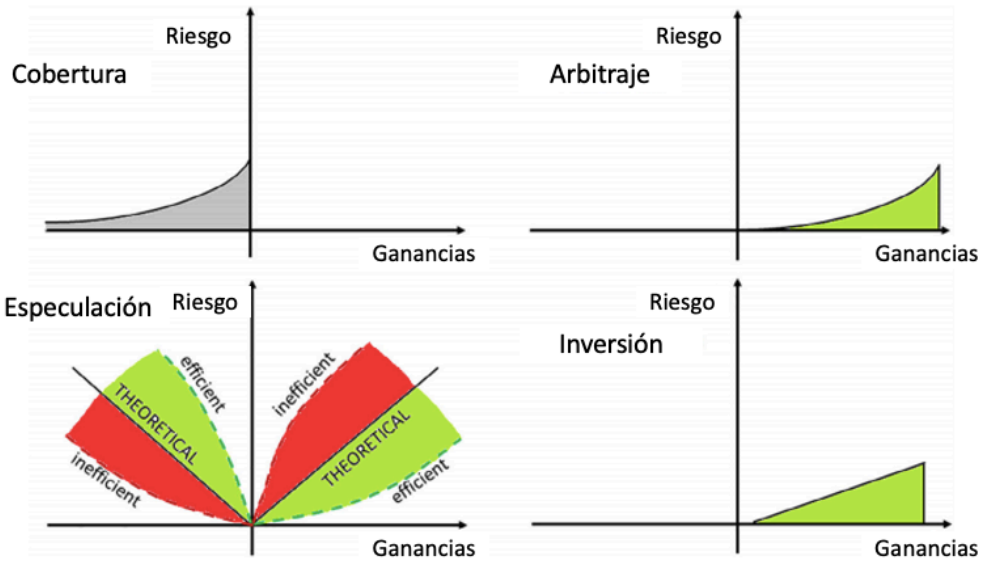

Figura 2 Ganancias vs riesgos en las 4 principales estrategias de mercado

La cobertura ocurre cuando el DM abre una posición de trading que compensa a otra posición tomada con antelación. Siguiendo este tipo de estrategia el inversionista busca protegerse de posibles caídas en los precios de los activos que conforman su portafolio.

El arbitraje toma ventaja de las diferencias de los precios de un activo entre mercados financieros diferentes en un mismo periodo de tiempo; lo que permite una ganancia libre de riesgo (Dybvig & Ross, 2003).

Las últimas dos categorías resultan ser muy similares, sus diferencias radican en la aversión al riesgo y la duración de las posiciones ejecutadas por el DM. En la inversión se supone que el DM realiza un análisis profundo sobre el posible comportamiento de un activo a lo largo del tiempo y determina el porcentaje de inversión a destinar por periodos extendidos, de esta forma logra beneficiarse de elementos subyacentes como las tasas de interés. En cambio, la especulación implica operar en lapsos de tiempo más reducidos, de tal forma que las ganancias son resultado de las fluctuaciones del valor de los activos en el mercado (Branco, 2017).

En un escenario de posiciones especulativas, alta incertidumbre e intervalos de operación bajos (días, horas, minutos, etc.), la tarea de asignación de capital por un inversionista humano se vuelve infactible. El agente de DRL propuesto en este trabajo logra desenvolverse como un especulador, en el escenario antes mencionado, ejecutando portafolios de inversión diferentes una vez por día en la apertura del mercado, mediante la realización de análisis técnico basado en los precios históricos e indicadores de activos de NYSE.

2.1.2 Análisis técnico

A la búsqueda de patrones en el comportamiento histórico de los mercados para predecir tendencias futuras se le conoce como análisis técnico.

Figura 3 Elementos del análisis técnico

Considerando la teoría del análisis técnico, se asume que toda la información relevante del mercado, para la toma de decisiones de trading, está reflejada en los precios de los instrumentos financieros, tomando en cuenta un intervalo en el tiempo (día, hora, minuto, etc.) estos serían; precio de apertura, de cierre, más alto, más bajo y volumen comercializado.

Aunado a los precios, los indicadores técnicos son una parte fundamental del análisis técnico. Dichos indicadores son valores computados a partir de las variables del mercado y agrupados en las siguientes categorías: tendencia, *momentum*, volatilidad e indicadores de volumen. En este trabajo se plantea el uso de algunos de los indicadores técnicos empleados por (Li, Zheng, & Zheng, 2019) para el entrenamiento de un agente de DRL.

2.1.3 Trading algorítmico

El sector financiero se ha visto beneficiado desde hace al menos cinco décadas por los avances tecnológicos y computacionales. Tareas como la comunicación con los agentes de bolsa, las transacciones de recursos, los pronósticos y la construcción de portafolios de inversión ahora pueden ser realizadas en lapsos de tiempo reducidos con muy poca intervención humana.

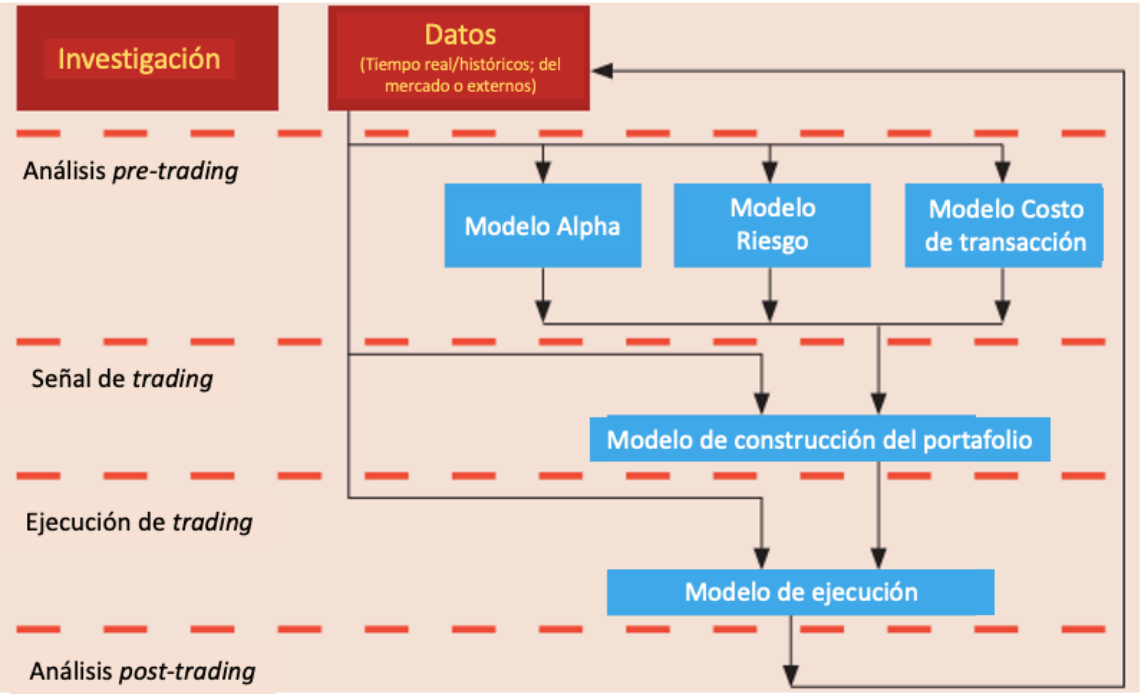

Figura 4 Componentes de un sistema de trading algorítmico

El trading algorítmico (AT) es definido como el uso de programas computarizados que de manera automática realizan las tareas del ciclo de comercialización; por ejemplo, la toma de decisiones de trading, envío de ordenes, gestión de ordenes, etc. Las etapas claves de un sistema de AT son: la investigación y recopilación de los datos, el análisis de pre-trading, la generación de señales de trading, la ejecución de dichas señales y el análisis post-trading. Los algoritmos empleados típicamente, determinan el momento adecuado, el precio, y la cantidad de valores a comercializar incluso simultáneamente en múltiples mercados (Treleaven, Galas, & Lalchand, 2013) (Hendershott, Jones, & Menkveld, 2011).

En este trabajo se ha desarrollado el método GA-DRL para las tres primeras etapas de un sistema de AT. (1) consiste en la recopilación y curado de datos, (2) construcción de un modelo MDP donde el espacio de estados esta conformado por observaciones de múltiples series de tiempo derivadas de activos de NYSE y BSE, el espacio de decisiones consiste en portafolios de inversión factibles y como función de recompensa se proponen dos versiones; el porcentaje de variación del portafolio de inversión en un intervalo de tiempo y el Ratio de Sharpe, (3) con motivo de evaluar la calidad y el rendimiento de la estrategia derivada del modelo MDP, la

etapa de generación de señales de trading involucra la aplicación de dicha estrategia en un periodo de prueba desconocido para el agente de DRL propuesto.

2.2 Métricas de inversión

Con el fin de conocer la calidad y rendimiento de las estrategias de inversión, los DMs emplean diversas métricas que les permiten evaluar la utilidad y el riesgo de sus portafolios. Dos métricas comúnmente usadas son la relación de retorno (ROR) y la ratio de Sharpe (SR).

Si bien estas métricas nos permitirán evaluar el desempeño del agente de inversión, se vuelven de gran importancia para el agente de DRL propuesto debido a su incorporación como funciones de recompensa/castigo para el proceso de aprendizaje.

2.2.1 Relación de retorno

La relación de retorno (ROR) es la ganancia o perdida de una inversión sobre un determinado periodo de tiempo, expresado como un porcentaje del monto inicial de la inversión.

$$ROR = \frac{Valor\ final\ del\ recurso - Valor\ inicial\ del\ recurso}{Valor\ inicial\ del\ recurso} \times 100 \qquad (2)$$

Un porcentaje positivo del ROR indica que el comportamiento del portafolio ejecutado por el inversionista ha generado ganancias. Por el contrario, un porcentaje negativo implica perdidas.

2.2.2 Ratio de Sharpe

La ratio de Sharpe (SR) es utilizada para calcular el retorno ajustado al riesgo (es decir, que tanto riesgo hay involucrado en producir cierta utilidad). Representa la diferencia en la utilidad de un portafolio financiero y el ROR de una inversión libre de riesgo, todo esto dividido por el riesgo del portafolio.

$$SR = \frac{Retorno\ promedio - Ratio\ libre\ de\ riesgo}{Desviación\ estándar\ del\ retorno} \qquad (3)$$

Esta métrica responde a la siguiente pregunta ¿el inversionista es capaz de conseguir un retorno por encima de un punto de referencia, pero con un riesgo más bajo? Por lo que, un portafolio de inversión con un Ratio de Sharpe mayor es preferible.

A continuación, se muestra la interpretación de los valores resultantes que puede tomar el SR (Sharpe, 1994):

- Si es negativo, el portafolio se ha comportado peor que un punto de referencia y la situación es mala: el portafolio tiene un rendimiento menor que una inversión libre de riesgo.
- Si se encuentra entre 0 y 1, el exceso de rendimiento del portafolio con respecto al punto de referencia es alcanzado debido a que se tomó un riesgo muy elevado.
- Si es mayor que 1, el retorno del portafolio es superior al punto de referencia para un riesgo apropiado. Es decir, la ganancia superior obtenida no está relacionada a la toma de un riesgo alto.

2.3 Aprendizaje por refuerzo

El aprendizaje por refuerzo (RL) es uno de los paradigmas de ML para el desarrollo de agentes artificiales involucrados en la toma de decisiones secuenciales y aprendizaje *online* en entornos con dinámica desconocida. De forma práctica puede ser visto como un método para el mapeo de situaciones del entorno (ambiente) a decisiones del agente, teniendo como objetivo maximizar una señal numérica de recompensa sobre un periodo de tiempo establecido. En RL no existen ordenes explicitas hacia el agente sobre que decisiones tomar, si no que por si mismo, mediante la prueba y error, este se encarga de descubrir cuales decisiones le traen el mayor beneficio a largo plazo (Sutton & Barto, 2018).

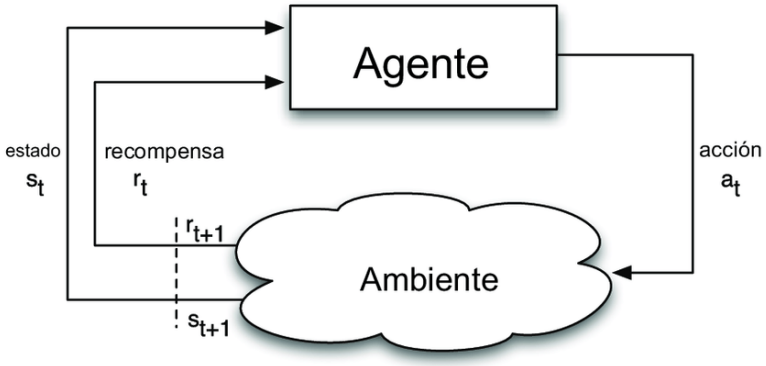

Figura 5 Escenario de RL; un agente interactuando con su entorno

Dos de las características más distintivas de RL, con respecto a los otros paradigmas de ML, son la recompensa desfasada y el aprendizaje mediante la experiencia de la prueba-error para determinar que decisiones ofrecen mayores beneficios al agente.

Los problemas de RL son formalizados como procesos de decisión de Markov (MDP), definidos por una tupla $M = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$; considerando pasos de tiempo discretos $t = 0, 1, 2, 3, \ldots$ en cada t el agente percibe una representación del estado del entorno $S_t \in \mathcal{S}$, donde $\mathcal{S}$ es el conjunto de posibles estados, a partir de la información recibida toma una decisión $A_t \in \mathcal{A}(S_t)$, donde $\mathcal{A}(S_t)$ es el conjunto de posibles decisiones disponibles en el estado S_t. La toma de decisión es realizada en base a una política π_t, donde $\pi_t(s)$ representa la decisión tomada si $S_t = s$. Para el siguiente $t + 1$, el agente recibe una recompensa R_{t+1} e información del nuevo estado S_{t+1} con motivo de la decisión tomada A_t. El componente $\mathcal{T}$, llamado matriz de probabilidades de transición, define la dinámica del sistema, para el problema abordado en este trabajo, $\mathcal{T}$ está determinada completamente por los datos disponibles del mercado y se desconoce su composición. Por último, γ es un factor de descuento que establece el peso dado a la recompensa recibida por el agente a través del tiempo.

En lo sucesivo, se mencionan los componentes principales de un entorno de RL (Sutton & Barto, 2018) ofreciendo ejemplos de implementación y el enfoque aplicado en esta tesis desde el punto de vista de trading algorítmico:

2.3.1 Agente

Es la entidad que, mediante la prueba-error y la orientación al cumplimiento de una meta, toma decisiones en el entorno para la construcción de estrategias de comportamiento que provean de beneficios a largo plazo. Los agentes de RL son diseñados con objetivos en especifico, mecanismos para la percepción de distintos aspectos del entorno y comportamientos o decisiones que influyen al entorno. Además, se asume que el agente es capaz, desde el inicio, de operar aún a pesar de la incertidumbre a la que se enfrente en el entorno.

Los ejemplos de agentes en entornos de RL son variados y comúnmente desarrollados para problemas en especifico. A continuación, se mencionan algunos (Sutton & Barto, 2018):

- En el caso de un agente para el movimiento de piezas en un juego de ajedrez; su objetivo constaría de ganar la partida. La percepción sobre el entorno estaría determinada por las posiciones de las piezas de ajedrez tanto propias como del oponente. El espacio de decisiones serían los posibles movimientos disponibles para cada pieza de acuerdo a su posición actual.
- Para un robot recolector de basura; su objetivo constaría de recolectar tantos objetos como sea posible. La percepción de su entorno podría estar determinada por sensores ópticos que permitan identificar los objetos considerados basura. Por último, el espacio de decisiones constaría de los movimientos de desplazamiento del robot (avanzar o girar) y la activación de un brazo o lampón de recolección.

El agente de trading algorítmico propuesto para este trabajo tiene como objetivo la construcción la generación de estrategias de inversión con ganancias superiores a la estrategia *Buy & Hold*. La percepción del entorno consiste en la observación de variables del mercado o indicadores técnicos y sus decisiones constan de los portafolios de inversión factibles a ejecutar.

2.3.2 Estado

El estado representa la condición actual del entorno, a partir de este se establece la decisión a tomar por el agente. La composición del estado determina el rendimiento del agente en circunstancias reales de interacción con el entorno, es por ello que la elección de los componentes del estado es de suma importancia a la hora de definir un problema en el paradigma de RL.

Al condicionar el estado siguiente S_{t+1} a la decisión A_t, es necesario que se contemple el contexto en el que se encuentra un agente en determinado momento. Retomando el ejemplo del ajedrez, para una cierta configuración del tablero se buscaría que el agente actuara con mayor cautela si los movimientos del juego han sido menores a cierto número establecido y de forma más agresiva en caso contrario. Para este caso se requeriría tener un registro completo sobre los estados, acciones y recompensas que el agente ha experimentado para poder tomar una decisión A_t en el estado actual S_t, lo cual se expresaría de la siguiente forma:

$$\Pr\{S_{t+1} = s', R_{t+1} = r \mid S_0, A_0, R_1, \dots, R_t, S_t, A_t\} \tag{4}$$

Sin embargo, se considera que, aún en sistemas complejos, no es necesario tener un registro de todos los eventos ocurridos hasta el momento t para la toma de decisiones y se asume que toda la información relevante para las decisiones futuras se encuentra contenida en el estado actual. A los estados con esta característica se les conocen como Markovianos o estados que cumplen la propiedad Markov. De tal que forma que es posible reducir la expresión 4 a:

$$\Pr\{S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a\} \tag{5}$$

Para nuestro caso, en donde el entorno es la representación de un mercado financiero, los estados son definidos por variables del mercado (precios y volumen de comercialización) e indicadores técnicos (transformaciones de los valores de las variables del mercado). En este trabajo asumimos que los estados del entorno cumplen la propiedad de Markov por lo que se pretende que sus componentes representen en gran medida toda la información relevante para la toma de decisiones por parte del agente.

2.3.3 Recompensa

La señal de recompensa define el objetivo a alcanzar en un problema de RL. Tal como se comentó al inicio de esta sección, en cada momento t el agente recibe una retroalimentación numérica por parte del entorno, el único objetivo del agente es maximizar el total de dicha señal; la cual recibe a lo largo del tiempo de interacción con el entorno. Por lo tanto, la señal de recompensa debe de ser diseñada para definir el beneficio que aportan los diferentes eventos a los que se ve expuesto el agente.

Nuestro entorno ha sido diseñado con las métricas ROR y SR como funciones de recompensa con el objetivo de calificar los portafolios de inversión a ejecutar.

2.3.4 Decisiones

La meta de un agente de RL es descubrir las decisiones mas apropiadas para las situaciones o estados que se presenten; es decir, una política π para la que se obtenga la mayor recompensa posible con respecto a un periodo de tiempo establecido. El agente recibirá, a través de la interacción con el entorno, una serie de recompensas $R_{t+1}, R_{t+2}, R_{t+3}, \ldots$ que buscará maximizar.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{T} \gamma^k R_{t+k+1} \tag{6}$$

Esta acumulación de recompensa, en un horizonte de tiempo, permite asignar valor a los estados mientras se sigue una política π, a esta función se le conoce como función de valor $V_\pi(s)$.

$$V_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid s = s_t] = \mathbb{E}_\pi\left[\sum_{k=0}^{T} \gamma^k R_{t+k+1} \mid s = s_t\right] \tag{7}$$

Por medio de $V_\pi(s)$ es posible considerar mas deseables a los estados de alto valor (teniendo en cuenta el beneficio a largo plazo) en lugar de solo acciones de alta recompensa (beneficio a corto plazo).

El espacio de decisiones, para el agente de trading desarrollado, esta constituido por portafolios de inversión factibles, representados como un vector donde cada elemento determina el porcentaje de capital a destinar a un activo en particular $a_t = w_t = [w_{1,t}, w_{2,t}, \dots, w_{N,t}]$; donde $a_t \in A \subseteq [0,1]^N, \forall t \geq 0 \; subjeto \; a \; \sum_{i=0}^{N} a_{i,t} = 1$.

2.3.5 Política

La política es una función para el mapeo de estados a decisiones; es decir, determina la actuación del agente en base a su percepción actual del entorno. En el algoritmo PPO la política óptima $\pi^*(s)$ es desarrollada de manera indirecta en base a la función $V_\pi(s)$ que es alimentada por el muestreo de las interacciones que el agente tiene con el entorno. Por lo que, al finalizar el proceso de aprendizaje se espera que los valores $V^*(s) \; \forall s \in \mathcal{S}$, definan la política de comportamiento deseada $V^*(s) = max_\pi \, V_\pi(s) \; \forall \; s \in \mathcal{S}$.

Para nuestro agente de trading algorítmico, la política determina, en cada momento y dado el estado actual del sistema, un portafolio de inversión a ejecutar.

2.3.6 Modelo del entorno

Es el componente de interacción directa con el agente; una representación suficiente de un entorno para el cual buscamos encontrar una estrategia de comportamiento (política) óptima. La utilidad del modelo radica en la retroalimentación ofrecida al tomar una decisión en el estado actual S_t; es decir, la obtención de un estado siguiente S_{t+1} al que transitar y la recompensa R_{t+1}.

Continuando con los ejemplos presentados en la sección previa:

- El modelo del entorno en un juego de ajedrez podría ser constituido por las piezas y sus posiciones dentro del tablero, ofreciendo retroalimentación de acuerdo a los términos de conclusión de una partida.
- En el caso del robot recolector de basura; el modelo del entorno estaría conformado por las imágenes obtenidas a partir de una cámara de video y la retroalimentación dada por la recolección exitosa de un objeto.

El grado de detalle del modelo del entorno a utilizar para entrenar a un agente de RL depende directamente de la información disponible del problema, por lo que entre más variables relevantes sean tomadas en cuenta, mayor será la utilidad de lo aprendido por el agente para conducirse en un entorno real. Además, el modelo del entorno, es un compendio de los componentes que se han venido describiendo hasta el momento (estados y función de recompensa) en el proceso de aprendizaje de un agente de DRL. En cuanto al diseño del modelo del entorno, es una de las actividades de más importancia al momento de plantear un problema de RL ya que a partir de él se desarrolla la estrategia de comportamiento que se busca emplear en el entorno real y de la cual se pretenden obtener resultados favorables (altos beneficios a largo plazo).

2.4 Aprendizaje profundo

El aprendizaje profundo (DL) forma parte del aprendizaje de representación ya que las máquinas de este tipo pueden ser alimentadas a partir de datos crudos y automáticamente descubrir representaciones necesarias para la clasificación o detección. DL permite a modelos computacionales que están compuestos de múltiples capas de procesamiento aprender representaciones de datos con múltiples niveles de abstracción. Mediante el uso del algoritmo *backpropagation*, DL descubre estructuras intrincadas en conjuntos de datos grandes indicando a la máquina como debería de cambiar sus parámetros internos usados para computar la representación en cada capa. Los métodos de DL permiten el aprendizaje de funciones complejas mediante la composición de módulos no-lineales que transforman la representación en cada capa para ser entregada a la capa siguiente. Una particularidad de DL es que las capas de características son definidas a partir de los datos de entrada y no por intervención humana (LeCun, Bengio, & Hinton, 2015).

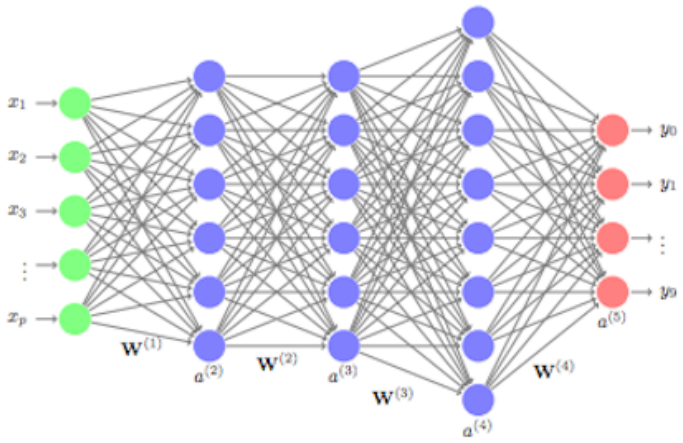

Figura 6 Ejemplificación de una red neuronal de DL

La implementación habitual de DL son las redes neuronales artificiales con múltiples capas cuya composición y funcionamiento se expone a continuación.

2.4.1 Redes neuronales artificiales

Las redes neuronales artificiales (ANN) son modelos matemáticos para la imitación del comportamiento de cerebros biológicos que tienen como objetivo la clasificación o identificación de representaciones de objetos. El elemento fundamental de una ANN es la neurona cuya función principal es la transformación de características de un objeto (datos de entrada) y la producción de una salida que identifique o califique de alguna forma a la instancia de entrada. Una ANN puede ser vista como un conjunto de neuronas conectadas entre si a lo largo de múltiples capas.

En 1958 Rosenblatt, siguiendo el modelo simplificado de una neurona biológica, desarrolla el perceptron (neurona artificial); un modelo/algoritmo capaz de aprender a discriminar y reconocer patrones perceptivos (Rosenblatt, 1960). Los componentes principales de esta neurona artificial se describen a continuación:

Entradas: representan características del objeto que se desea clasificar, detectar o valorar.

Pesos: son parámetros a combinar con las entradas para producir una salida, su valor es dinámico conforme la neurona es alimentada por más datos.

Función de activación: junto a una función de agrupación, la función de activación transforma, comúnmente de forma no-lineal, las entradas y pesos en un valor normalizado.

Salida: mediante una regla sencilla de decisión, el producto de la función de activación es transformado a una categoría, monto o valor.

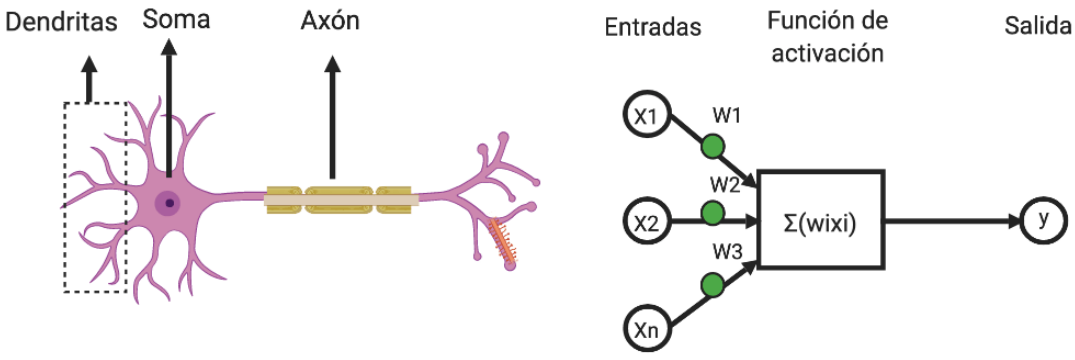

Figura 7 Modelo simplificado de una neurona biológica (izquierda) y neurona artificial (derecha)

2.5 Aprendizaje por refuerzo profundo

El aprendizaje por refuerzo profundo (DRL) es el resultado de la combinación tanto de técnicas de RL como de DL para problemas donde el posible espacio de estados es considerado demasiado grande para que sus valores sean almacenados de forma tabular en memoria como habitualmente se hace en los algoritmos básicos de RL como Q-learning o SARSA. DRL se apoya en el concepto de aproximadores de funciones, como lo pueden ser las DNN, las cuales son alimentados con entradas del tipo estado o estado-acción y producen una salida numérica interpretada como el valor o directamente la decisión tomada por el agente (Zai & B., 2020)

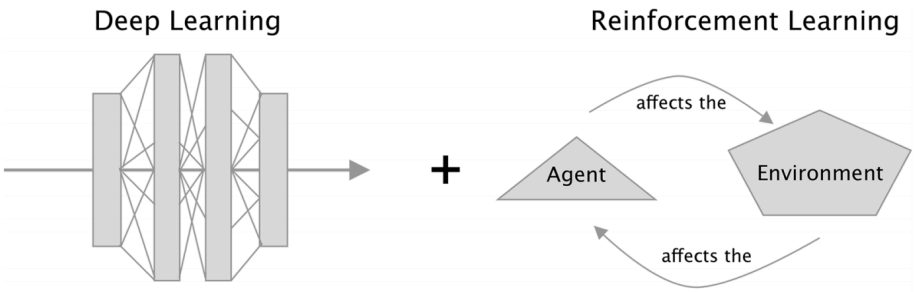

Figura 8 Aprendizaje profundo + aprendizaje por refuerzo

Dejando de lado el modo tabular de almacenar todos los estados de nuestro entorno, lo cual puede llegar a ser imposible en muchos problemas reales, es posible entrenar modelos de DL con muestreos de estados a partir de los cuales podemos predecir que tan valiosos son con respecto al objetivo definido para nuestro agente de DRL.

Algunos modelos de DL como las DNN utilizan coeficientes/pesos para poder aproximar la función que relaciona entradas con salidas y el objetivo es encontrar los coeficientes/pesos más apropiados mediante el ajuste iterativo de los mismos empleando gradientes que produzcan un valor menor para una función de perdida.

3 Trabajos relacionados

Este capítulo incluye un resumen de los trabajos recientes sobre DRL aplicado a trading algorítmico. En la Tabla 1 se exponen los puntos en común y diferencias entre los proyectos.

En (Li, Zheng, & Zheng, 2019) se propone un agente de trading, basado en DRL, para trading autónomo y generación ganancias en los mercados de acciones y futuros, obteniendo retornos anualizados superiores a la estrategia *Buy & Hold* y valores de la ratio de Sharpe mayores en comparación con las estrategias de RL en su versión básica. Se realizaron modificaciones a los algoritmos *deep Q-network* (DQN) y *asyncronous advantage actor critic* (A3C) con el fin de ser eficientes en el proceso de comercialización. Para lidiar con la extracción de características relevantes del mercado y el formato de los datos de entrada (series de tiempo financieras) se emplearon stacked denoising autoencoders (SDAEs) y redes neuronales de tipo long short-term memory (LSTM) como parte de la función de aproximación. En el marco del MDP a resolver, los estados son compuestos por; variables del mercado (precio de apertura, cierre, mas alto, mas bajo y volumen comercializado), indicadores técnicos (MACD, MA, EMA, ATR y ROC) y variables privadas (efectivo disponible y ratio de Sharpe previo). El espacio de decisiones $\{-n, -n+1, \dots, 0, n-1, n\}$ representa el número de acciones a vender (positivo) o comprar (negativo). Por último, la función de recompensa está definida como $r_t = \Delta c p_{t-1} - (\alpha + \beta)|\Delta p|$ donde α es el porcentaje del costo de la transacción, β es el porcentaje de *slippage*, Δc es el cambio de precio de cierre y Δp es el cambio de posición del agente inversionista.

En el trabajo de (Ponomarev, Oseledets, & Cichocki, 2019) se presenta un sistema para trading de un instrumento financiero individual, obteniendo una estrategia de inversión para el índice de futuros RTS con ganancias anuales del 66%. El algoritmo de RL propuesto es del tipo *asyncronous advantage actor critic* (A3C) con experimentos en diversas arquitecturas de redes neuronales. Con respecto al modelo MDP, el estado es un agregado de las ofertas actuales. El espacio de decisiones $\{-1, 0, 1\}$ representa: mantener una unidad del instrumento financiero (1), mantenerse neutral (0) y tomar prestada una unidad del instrumento (-1). Por último, la función de recompensa r_t es calculada a partir de la variación entre la ganancia de los activos poseídos en el tiempo t y $t-1$.

Utilizando el algoritmo *Deep deterministic policy gradient* (DDPG), (Conegundes & Pereira, 2020) resuelven el problema de asignación continua de recursos financieros; definiendo los porcentajes del capital disponible a destinar a un portafolio de activos en múltiples periodos de tiempo, dando como resultado una estrategia con retorno acumulado, sobre 3 años, del 311% y un *drawdown* máximo promedio anual de alrededor de 19%. Con respecto al modelo MDP, el estado es un vector de los precios relativos de apertura y cierre por una ventana de tiempo de longitud W. Una decisión está definida por el vector $w_t = [w_{0,t}, w_{1,t}, \dots, w_{N,t}]$ que representa el capital (en porcentaje) destinado a cada activo dentro del portafolio. Para finalizar, la función de recompensa se establece como $r_t = p_{t-1} y_{t-1} * w_{t-1}$ siendo p_{t-1} el valor inicial del portafolio, y_{t-1} el vector de precios relativos y w_{t-1} el portafolio construido.

En el trabajo de (Lei, Zhang, Li, Yang, & Shen, 2020) se desarrolló un agente de RL para el trading con capacidad de seleccionar, de manera adaptativa, características de los datos de entrada y ponderarlas de acuerdo a la temporalidad, obteniendo estrategias para acciones en 11 sectores diferentes con un retorno anualizado promedio al menos 30% por encima de las estrategias *Buy & Hold* y *Deep direct reinforcement*. En el marco del MDP a resolver, los estados son compuestos por; una ventana de tiempo de longitud W de los indicadores seleccionados y ponderados. El espacio de decisiones $\{1, 0, -1\}$ representa la posición a tomar en determinado momento; $\{long, neutral, short\}$. Por último, la función de recompensa se establece como $r_t = \delta_t z_t - c|\delta_t - \delta_{t-1}|$ donde c es la comisión a pagar por la transacción, z_t el beneficio con respecto al movimiento previo y δ_t una transformación trigonométrica de la acción tomada por el agente y el estado actual del entorno.

En el trabajo de (Yuan, Wen, & Yang, 2020) se propone un *framework* de aprendizaje por refuerzo basado en ampliación de datos (DARL) para el trading diario de recursos financieros individuales, obteniendo retornos anualizados y ratios de Sharpe superiores con el algoritmo *Proximal policy optimization* (PPO), en comparación con la estrategia *Buy & Hold*, mientras que con DQN y *Soft actor critic* (SAC) solo ratios de Sharpe mayores en todos los casos. Con respecto al modelo MDP, el estado está compuesto por N barras con variables del mercado (precio de apertura, cierre, mas alto, mas bajo y volumen comercializado), un indicador del momento de compra de la acción y la ganancia o perdida de la posición actual. El espacio de decisiones está conformado de la siguiente forma: {comprar una acción, mantenerse, vender

una acción}. Para finalizar, la función de recompensa está definida como la ganancia o perdida obtenida al vender una acción.

A continuación, las características de los trabajos relacionados.

Referencia	Algoritmo de DRL	Prese-lección automática de activos	Serie de tiempo empleadas	Trading de portafolio	Métricas de Evaluación
(Li, Zheng, & Zheng, 2019)	DQN y A3C DNN LSTM (selección de características)		Acciones de futuros y valores (índice e individuales) Intervalos de 1 minuto		Retorno anualizado y ratio de Sharpe
(Ponomarev, Oseledets, & Cichocki, 2019)	A3C LSTM		Futuros (índice) Intervalos de 1 minuto		Retorno anualizado + comisión, ratio de Sharpe y ganancia promedio por transacción
(Conegundes & Pereira, 2020)	DDPG LSTM		Valores Intervalos de 1 día	✓	Retorno anualizado, retorno acumulado (3 años) y d*rawdown* promedio anualizado
(Lei, Zhang, Li, Yang, & Shen, 2020)	PG LSTM y GRU		Valores Intervalos de 1 día		Curva de ganancias, ganancias totales, ROR anualizado, ratio de Sharpe y

					número de transacciones realizadas
(Yuan, Wen, & Yang, 2020)	PPO, DQN y SAC DNN		Valores Intervalos de 1 día		Retorno anualizado y ratio de Sharpe
Este trabajo	AC MLP	✓ (GenPo-Sharpe)	Valores de NYSE Intervalos de 1 día	✓	Retorno anualizado, ratio de Sharpe, volatilidad, drawdown máximo.

Tabla 1 Trabajos relacionados sobre DRL para AT

4 Metodología propuesta

Este trabajo propone la metodología GA-DRL para la generación de estrategias de inversión para la gestión activa de portafolios de acciones de la NYSE. GA-DRL es una combinación de un método para selección de portafolios de inversión (GenPo-Sharpe) y un agente de DRL para la asignación continua de recursos financieros. El objetivo de este método es obtener ganancias superiores a las estrategias *Buy & Hold* también conocidas como gestión pasiva de portfolios.

En las figuras 9 y 10 se observan las fases de entrenamiento y prueba que describen de manera general la metodología:

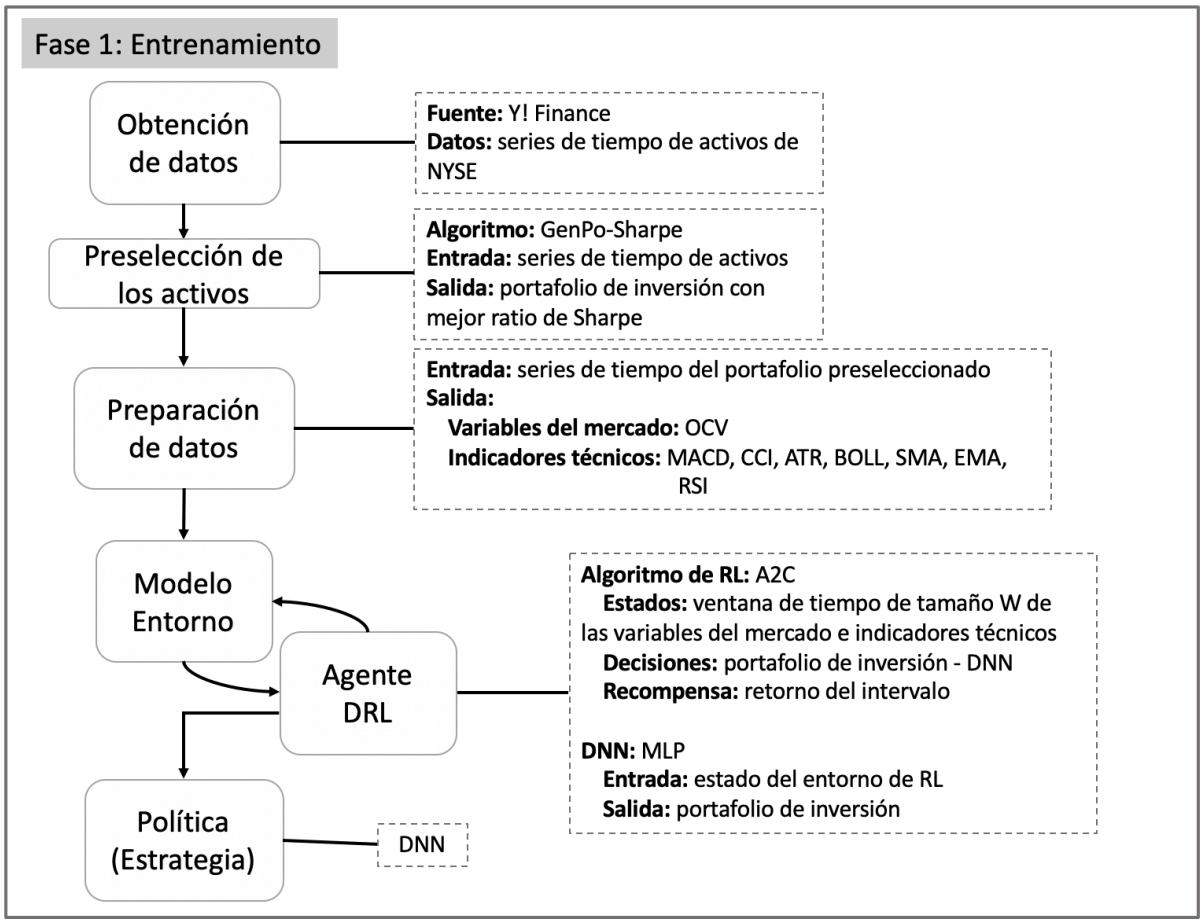

Figura 9 Fase de entrenamiento del método GA-DRL para AT.

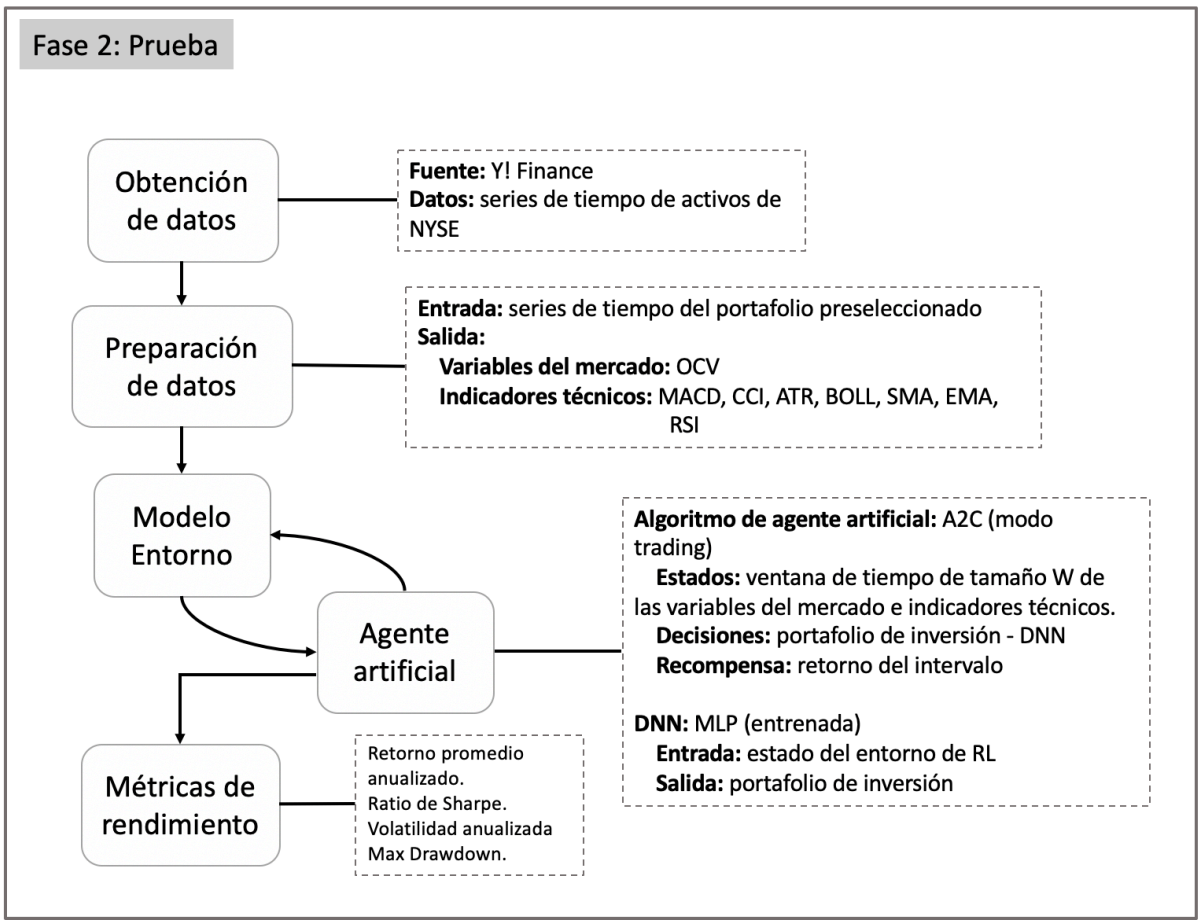

Figura 10 Fase de prueba del método GA-DRL para AT.

4.1 Selección de mercado financiero

Debido al incremento en el volumen de comercialización en el mercado de valores durante el año 2020, se planteó analizar las principales cinco bolsas de valores del mundo con el objetivo de seleccionar aquella que resultara mayormente atractiva para los inversionistas y fuera representativa de un mercado desarrollado, teniendo como principal criterio su capitalización de mercado.

Puesto	Bolsa de valores	Región	Capitalización de mercado (BDD)
1	New York Stock Exchange	Estados Unidos de América	25,62
2	Nasdaq	Estados Unidos de América	19,51
3	Hong Kong Exchanges	Hong Kong	6,76
4	Shanghai Stock Exchange	China	6,56
5	Japan Exchange Group	Japón	6,54

Tabla 2 Las cinco bolsas de valores mas grandes, por capitalización de mercado, a Febrero de 2021

(Statista Research Department, 2021)

De acuerdo al reporte de (Statista Research Department, 2021), la bolsa de Nueva York (NYSE) presenta la mayor capitalización (25,62 BDD) del mercado estando más de 6 billones de dólares por encima del segundo puesto el cual pertenece a Nasdaq.

4.2 Adaptación del algoritmo GenPo-Sharpe

El algoritmo GenPo-Sharpe (Frausto, y otros, 2020) es un algoritmo de tipo genético propuesto para la construcción de portafolios de inversión donde se tienen tres objetivos (1) maximización del valor esperado del portafolio, (2) minimización del riesgo del portafolio y (3) cumplimiento de las preferencias del DM sobre el número de activos a incluir en el portafolio.

GenPo-Sharpe ha sido puesto a prueba con activos de la bolsa Mexicana de Valores (acciones comunes y Fibras) y sometido a competición contra otros métodos de construcción de portafolios basados en el modelo de Markowitz, obteniendo rendimientos promedio superiores al método de programación cuadrática (QP) (Frausto, y otros, 2020).

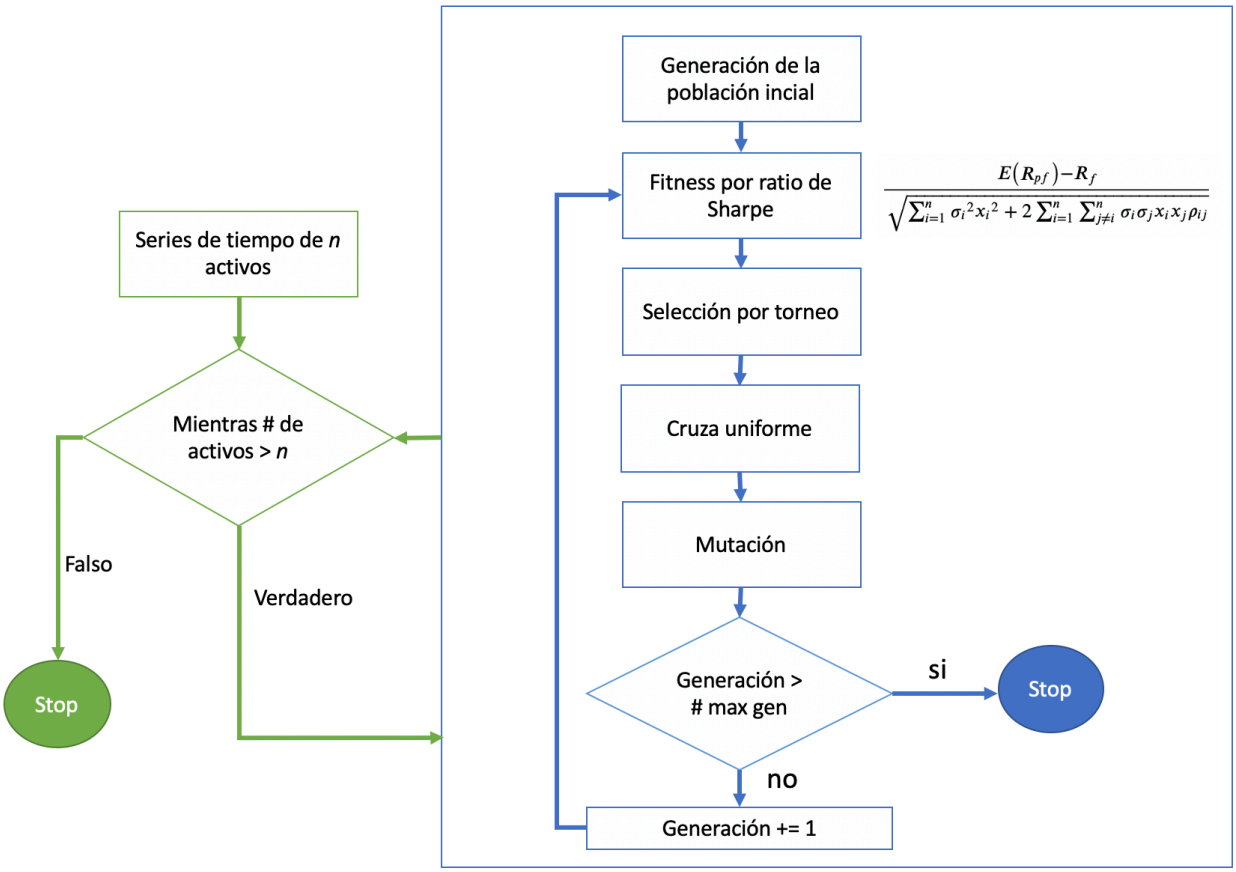

Figura 11 Algoritmo GenPo-Sharpe con adaptación para incrementar la eficiencia al trabajar con sectores financieros completos

basado en (Frausto, y otros, 2020)

En este trabajo de tesis, GenPo-Sharpe funge como pre-seleccionador de activos, por una parte, reduciendo el número de activos que componen a los sectores de Energía, Salud y Tecnología de NYSE, a uno determinado por las preferencias del inversionista y por otra parte maximizando la Ratio de Sharpe que, como ya se ha señalado, contempla el valor esperado y riesgo de un portafolio de inversiones.

En la figura 11 se puede observar, en el recuadro azul, la implementación original del algoritmo y en el recuadro verde las adaptaciones necesarias para hacerlo eficiente a la hora de crear portafolios con número de activos reducido cuando es alimentado por todos los activos de un sector financiero.

4.3 Establecimiento de componentes de MDP

Los componentes del MDP nos permiten modelar tres elementos de importancia en nuestro problema (1) el entorno de interacción (mercado de valores) que se encuentra caracterizado por estados que a su vez están definidos por variables del mercado e indicadores técnicos que se discuten a continuación (2) el agente inversionista (función valor y política) encargado de la toma decisiones que afectan al entorno y (3) una función de recompensa que asigna un valor numérico a las decisiones tomadas por el agente en cada instante de tiempo.

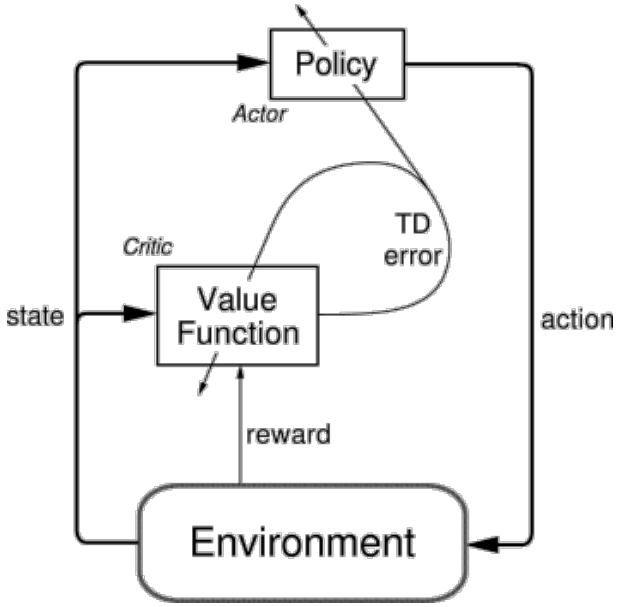

Figura 12 Escenario de RL para AT

4.3.1 Estados

Las diferentes configuraciones propuestas para la definición de los estados del entorno fueron inidentificadas por el análisis de los trabajos de (Dash & Dash, 2016), (W., Yue, & Rao, 2017), (Vanstone & Finnie, 2009) donde se trata el problema de pronóstico y toma de decisiones empleando técnicas de ML en series de tiempo financieras. El criterio para la selección de las variables e indicadores técnicos consistió en la existencia de las mismas en los tres trabajos citados.

Nombre	Definición
Variables del mercado	
Open, Close	Precio de apertura y cierre diario.
Volume	Volumen comercializado diario.
Indicadores técnicos	
SMA_[5,10]	Media móvil simple de 5 y 10 periodos.
EMA_[5,10]	Media móvil exponencial de 5 y 10 periodos.
MACD	Media móvil de convergencia y divergencia: muestra características de seguimiento de tendencia y *momentum*.
RSI_14	Índice de fuerza relativa: mide la magnitud de los cambios recientes en el precio para evaluar condiciones de sobre compra o sobre venta.
BOLL_20	Bandas de *Bollinger*: permiten aplicar la teoría de Reversión a la media para la que se asume que todo precio, a pesar de sus fluctuaciones, eventualmente regresa a su promedio.
CCI	Índice de canal de *commodities*: ayuda a identificar el inicio y final de una tendencia.
ATR	Rango promedio real: mide la volatilidad del precio

Tabla 3 Variables de mercado e indicadores técnicos que componen a los estados del mercado de valores

4.3.2 Decisiones

Las decisiones A_t del agente de trading están definidas como los portafolios de inversión a ejecutar en cada instante t. Cada portafolio es representado como un vector w_t donde los elementos del mismo hacen referencia al porcentaje del capital destinado a cada uno de los N activos en el día t.

$$SR = a_t = w_t = [w_{1,t}, w_{2,t}, \dots, w_{N,t}] \quad (4)$$

Dado que nuestro agente no tiene la capacidad de realizar operaciones de tipo *short*, los pesos destinados a cada activo son solo positivos:

$$a_t \in A \subseteq [0,1]^N, \forall t \geq 0 \; subjeto \; a \sum_{i=0}^{N} a_{i,t} = 1 \qquad (5)$$

4.3.3 Función de recompensa

Cómo funciones de recompensa para calificar las decisiones (portafolio de inversión) tomadas por el agente se han propuesto las métricas de evaluación presentadas en la sección 1.4:

Nombre	Fórmula
Relación de Retorno	$\frac{Valor\ final\ del\ recurso - Valor\ inicial\ del\ recurso}{Valor\ inicial\ del\ recurso} \times 100$
Ratio de Sharpe	$\frac{Retorno\ promedio - Ratio\ libre\ de\ riesgo}{Riesgo\ del\ portafolio}$

Tabla 4 Funciones de recompensa a implementar en el modelo del mercado de valores

La implementación de ambas funciones dio como resultado un total de ocho combinaciones de modelos del entorno para experimentar.

4.4 Algoritmo de DRL

La arquitectura del método de RL seleccionada en este trabajo, para la implementación del agente de trading, es de tipo Actor-Crítico, particularmente la implementación *Proximal Policy Optimization* (PPO) (Schulman, 2017) reportada en la literatura como uno de los mejores métodos para la solución de problemas de control donde el espacio de decisiones del agente está definido como un vector de elementos continuos y el espacio de estados del entorno es multivariable. (Mnih, y otros, 2016).

En la figura 13 se presenta el pseudocódigo de un algoritmo tipo Actor-Crítico. Dado que las decisiones que toma el agente, propuesto en este trabajo, son de tipo multivariable (vectores que representan un portafolio de inversión), en la línea 6 se observa el muestreo de una función con distribución Gaussiana construida a partir de la red actor $\pi^{\theta}(s)$. En las líneas 9, 10 y 11 se indican los cálculos de las funciones de perdida que se buscan minimizar en el entrenamiento de la red *critic* y *actor*.

Algorithm Actor-Critic

1. *Randomly initialize critic network* $V_{\pi}^{U}(s)$ *and actor network* $\pi^{\theta}(s)$ *with weights* U *and* θ
2. *Initialize environment E*
3. **for** *episode = 1, M* **do**
4. *Receive initial observation state* s_0 *from E*
5. **for** *t=0, T* **do**
6. *Sample action* $a_t \sim \pi^{\theta}(a|s_t, \theta, \Sigma)$ *according to current policy*
7. *Execute action* a_t *and observe reward* r *and next state* s_{t+1} *from E*
8. *Set TD target* $y_t = r + \gamma \cdot V_{\pi}^{U}(s_{t+1})$
9. *Update critic by minimizing loss:* $\delta_t = \left(y_t - V_{\pi}^{U}(s_t)\right)^2$
10. *Update actor policy by minimizing loss:*
11. $Loss = -log\left(\quad \mathbb{P}(a_t|\mathcal{N}(\mu, \Sigma))\quad\right) \cdot \delta_t$
12. *Update* $s_t \leftarrow s_{t+1}$
13. **end for**
14. **end for**

Figura 13 Algoritmo Actor-Crítico para trading algorítmico

A continuación, se muestran los diagramas de las redes neuronales implementadas para cumplir la función de política (actor) y función de valor (*critic*) basadas en el trabajo de (Schulman, 2017):

La red neuronal que cumple la función de política (figura 14) está compuesta por una capa de entrada x que define la representación actual del estado del entorno y dos capas internas con 64 neuronas cada una, su objetivo es estimar los valores de μ que interpretamos como un vector de las medias de las distribuciones normales, correspondientes a cada uno de los activos que conforma nuestro portafolio de inversión, este vector de distribuciones es empleado para el muestro y obtención de m variables $N(\mu, \Sigma)$ que son ajustadas a las restricciones de nuestro

problema mediante una capa de tipo *softmax* y dan como resultado el vector de decisión a_t (portafolio de inversión).

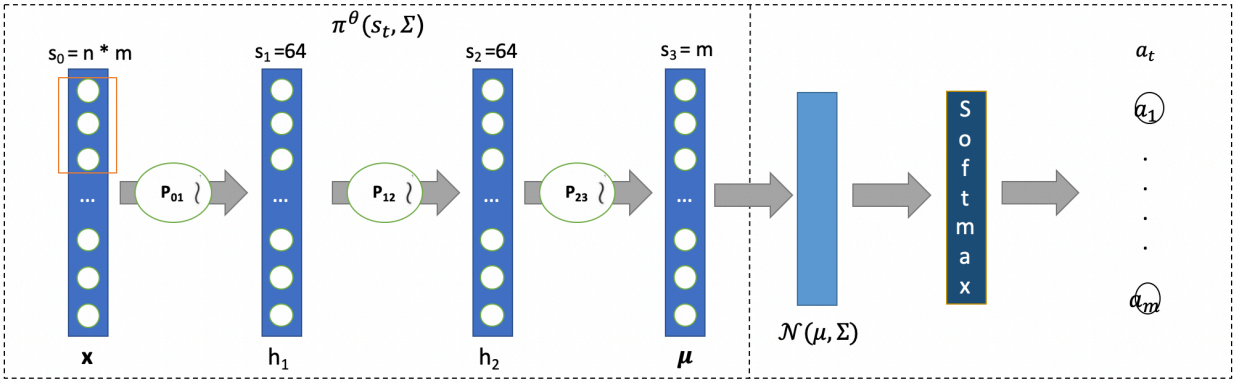

Figura 14 Red neuronal Actor

Donde:

x: vector de características (n indicadores técnicos y variables de mercado * m activos).

μ: vector de medias de distribuciones normales.

s_a: numero de neuronas.

P_{ab}: transformación lineal de las salidas de la capa a que alimentan a la capa b.

$\wr$: Transformación no-lineal del resultado P_{ab}: (función de activación $Tanh$).

Σ: Matriz de covarianzas.

a_t: Vector decisión.

La red neuronal que cumple la función de *critic* (figura 15) está compuesta por una capa de entrada x que define la representación actual del estado del entorno y dos capas internas con 64 neuronas cada una, su objetivo es estimar el valor del estado actual S_t utilizado para calificar y modificar el comportamiento del agente (política) a lo largo de sus interacciones con el entorno, tal y como se presentó en el pseudocódigo del algoritmo de tipo Actor-Crítico.

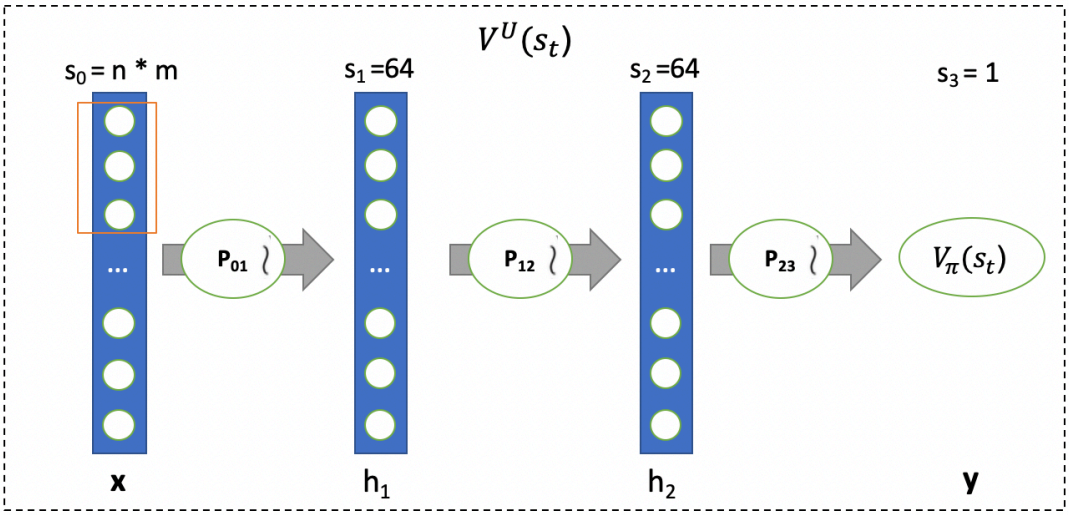

Figura 15 Red neuronal Critic

Donde:

x: vector de características (n indicadores técnicos y variables de mercado * m activos).

y: estimación de la valoración del estado actual s_t

s_a: numero de neuronas.

P_{ab}: transformación lineal de las salidas de la capa a que alimentan a la capa b.

$\wr$: Transformación no-lineal del resultado P_{ab}: (función de activación $Tanh$).

5 Experimentación y resultados

Los experimentos desarrollados en este trabajo de tesis consistieron en el cumplimiento de dos objetivos individuales; (1) la evaluación del método GA-DRL propuesto sobre activos de la NYSE y su comparación contra estrategias de inversión *Buy & Hold*, (2) la evaluación autónoma del agente inversionista de DRL con portafolios de activos pertenecientes a la BSE y su comparación contra el índice Ibovespa y métodos de DRL para la asignación continua de recursos.

Las métricas empleadas para medir la eficiencia de las estrategias son: (1) retorno acumulado y retorno anualizado que nos permiten conocer el cambio positivo/negativo en términos porcentuales del valor del capital inicial, (2) volatilidad anualizada cuya función es conocer el riesgo asociado a la variación del capital inicial siguiendo una estrategia en particular, (3) Ratio de Sharpe es un valor derivado del rendimiento del capital y el riesgo asociado a la estrategia utilizada, y por último, (4) drawdown máximo que refleja la disminución porcentual del capital más pronunciada previo a una recuperación en un periodo determinado.

5.1 Método GA-DRL

5.1.1 Descripción de los datos y diseño del experimento

El método GA-DRL empleado en esta experimentación consta de una primera etapa para la selección de un portafolio de activos financieros con un alto valor de ratio de Sharpe utilizando el algoritmo GenPo-Sharpe y una segunda etapa de asignación diaria de capital a los activos del portafolio seleccionado mediante el uso de un agente inversionista de DRL.

Los activos empleados para la experimentación computacional se derivan de los sectores Tecnología, Salud y Energía de la NYSE, existiendo en promedio 250 activos para cada sector a fecha de consulta de octubre del año 2020. El algoritmo GenPo-Sharpe se ejecutó con la preferencia de generar portafolios de máximo 10 activos y con precios de cierre del periodo del 1 de enero de 2020 al 31 de agosto de 2020. Los portafolios resultantes se muestran a continuación:

Tecnología	Salud	Energía
PLTR	ELAT	PDS
DELL	OSH	WLL
GLW	AMWL	PBA
LSPD	INSP	ALIN-PE
ST	CTLT	EURN
VNT	BIO	XOM
SAIC	TARO	TPL
YALA	BSX	DLNG-PA
CLGX	AMN	OVV
--------	TEVA	EPD

Tabla 5 Portafolios de inversión generados por GenPo-Sharpe

El agente inversionista de DRL fue entrenado en ocho diferentes variantes del entorno, en el periodo del 1 de octubre de 2020 al 30 de abril de 2021, derivadas de combinaciones de variables del mercado, indicadores técnicos y funciones de recompensa como se muestra en la siguiente tabla.

Variante	Variables	Indicadores técnicos	F. recompensa
E1	Open, Close, Volume	sma_[5,10], ema_[5,10], macd, rsi_14, cci, atr, boll_20 (1)	Sharpe Ratio (SR)
E2			Valor portafolio (PV)
E3		sma_[20], rsi_14, boll_20	SR
E4			PV
E5	Volume	sma_[5,10], ema_[5,10], macd, rsi_14, cci, atr, boll_20	SR
E6			PV
E7		sma_[20], rsi_14, boll_20	SR
E8			PV

Tabla 6 Variantes de modelo del mercado para agente de DRL

El periodo de evaluación, para las ocho estrategias de trading del agente de DRL que surgen a partir de las variantes de los modelos del mercado más dos estrategias de tipo Buy & Hold (GenPo-Sharpe, Equal Weighted) y una estrategia totalmente aleatoria, fueron evaluadas en el periodo del 1 de mayo de 2021 al 30 de junio de 2021.

5.1.2 Resultados

El primer experimento consiste en la comercialización de nueve activos del sector Tecnología de NYSE. A continuación, se presentan los resultados acordes a las métricas expuestas al inicio del capitulo:

	Retorno anualizado %	**Retorno acumulado %**	**Volatilidad anualizada %**	**Ratio de Sharpe**	**Drawdown máximo %**
GenPo-Sharpe	33.68	4.72	21.01	1.49	-6.71
Equal Weighted	32.00	4.50	22.00	1.37	-6.79
Random	36.96	5.12	21.81	1.55	-6.81
E1	81.48	9.92	26.87	2.35	-7.49
E2	127.80	13.96	30.56	2.85	-5.35
E3	43.90	5.95	25.46	1.55	-5.62
E4	56.53	7.37	28.46	1.71	-8.48
E5	118.40	13.20	32.22	2.58	-7.73
E6	61.23	7.88	24.89	2.04	-6.02
E7	92.28	10.94	25.93	2.65	-6.46
E8	31.61	4.46	26.61	1.16	-9.52

Tabla 7 Resultados de estrategias de trading para un portafolio de activos pertenecientes al sector Tecnología de NYSE en el periodo del 1 mayo de 2021 al 30 de junio de 2021

En la tabla 7 y resaltado en color amarillo podemos apreciar los mejores resultados globales de acuerdo a cada métrica evaluada. Siendo, para este caso, la estrategia del agente de DRL con variante del modelo del entorno E2, propuesta como método de esta tesis, la ganadora en cuatro del total de cinco métricas y dando un rendimiento al final del periodo de 13.96% con respecto al capital inicial.

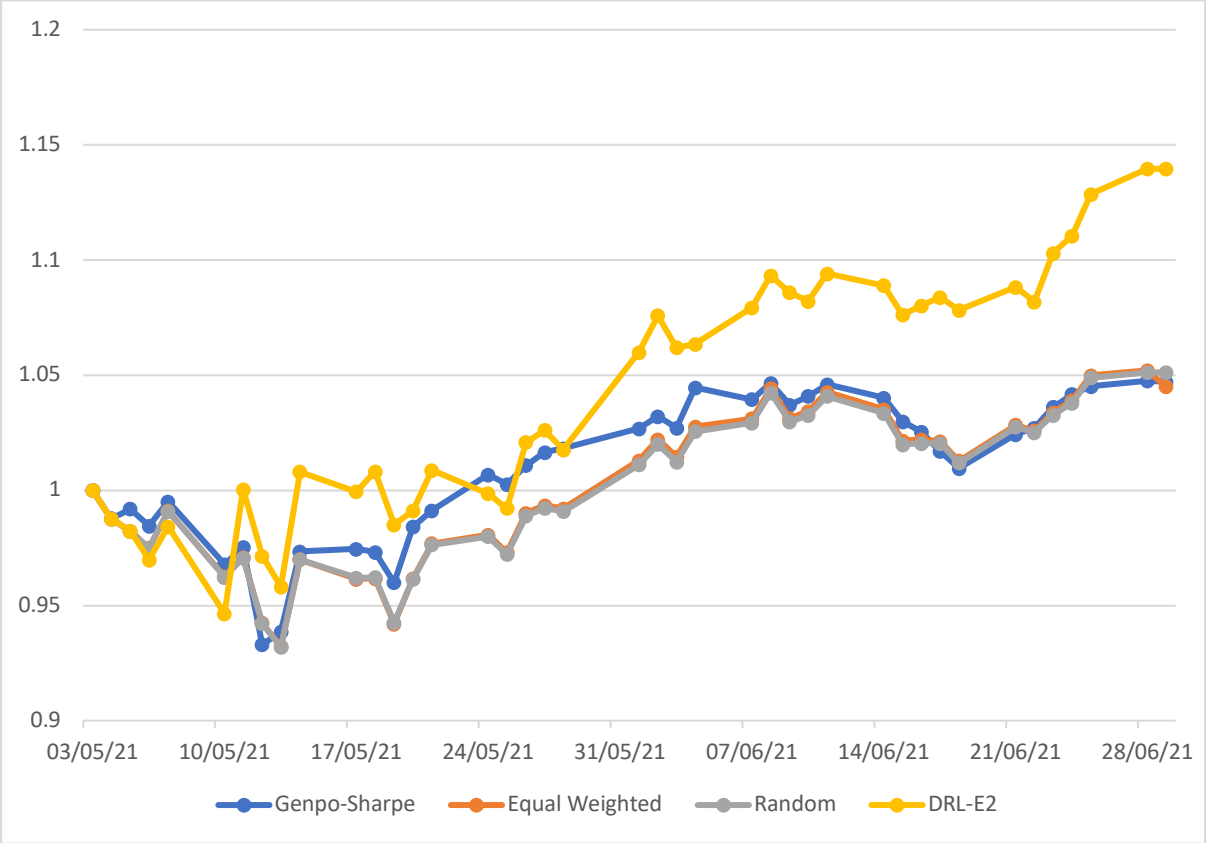

Figura 16 Rendimiento de estrategias para un portafolio de activos del sector Tecnología de NYSE en el periodo del 1 mayo de 2021 al 30 de junio de 2021

En la figura 16 podemos observar un rendimiento del capital acumulado del 172% y 195%, del agente de DRL con la versión del modelo del entorno E2 propuesto en esta tesis, por encima de las estrategias Random y GenPo-Sharpe respectivamente. Además, la preferencia de elección de esta estrategia DRL-E2 se vuelve más solida si consideramos que, su Ratio de Sharpe es 83% superior y su máximo drawdown 21.5% menor con respecto a la mejor de las tres estrategias de comparación.

Este experimento consiste en la comercialización de diez activos del sector Salud de NYSE. A continuación, en la tabla 8 se presentan los resultados:

	Retorno anualizado %	**Retorno acumulado %**	**Volatilidad anualizada %**	**Ratio de Sharpe**	**Drawdown máximo %**
GenPo-Sharpe	23.63	3.42	21.94	1.07	-7.20
Equal Weighted	-2.94	-0.47	19.30	-0.06	-8.82
Random	2.76	0.43	19.13	0.24	-8.93
E1	25.04	3.61	15.99	1.48	-5.00
E2	1.44	0.23	23.53	0.18	-12.33
E3	-14.61	-2.48	24.99	-0.51	-13.27
E4	15.54	2.32	14.57	1.06	-5.17
E5	-32.22	-5.99	24.53	-1.46	-13.03
E6	-4.79	-0.78	21.56	-0.12	-11.30
E7	22.37	3.26	20.05	1.10	-7.88
E8	-6.53	-1.07	26.69	-0.12	-13.76

Tabla 8 Resultados de estrategias de trading para un portafolio de activos pertenecientes al sector Salud de NYSE

Resaltado en color amarillo podemos apreciar los mejores resultados globales de acuerdo a cada métrica evaluada. Siendo, para este caso, la estrategia del agente de DRL con variante del modelo del entorno E1, propuesta como método de esta tesis, la ganadora en las cinco métricas evaluadas y dando un rendimiento al final del periodo de 3.61% con respecto al capital inicial.

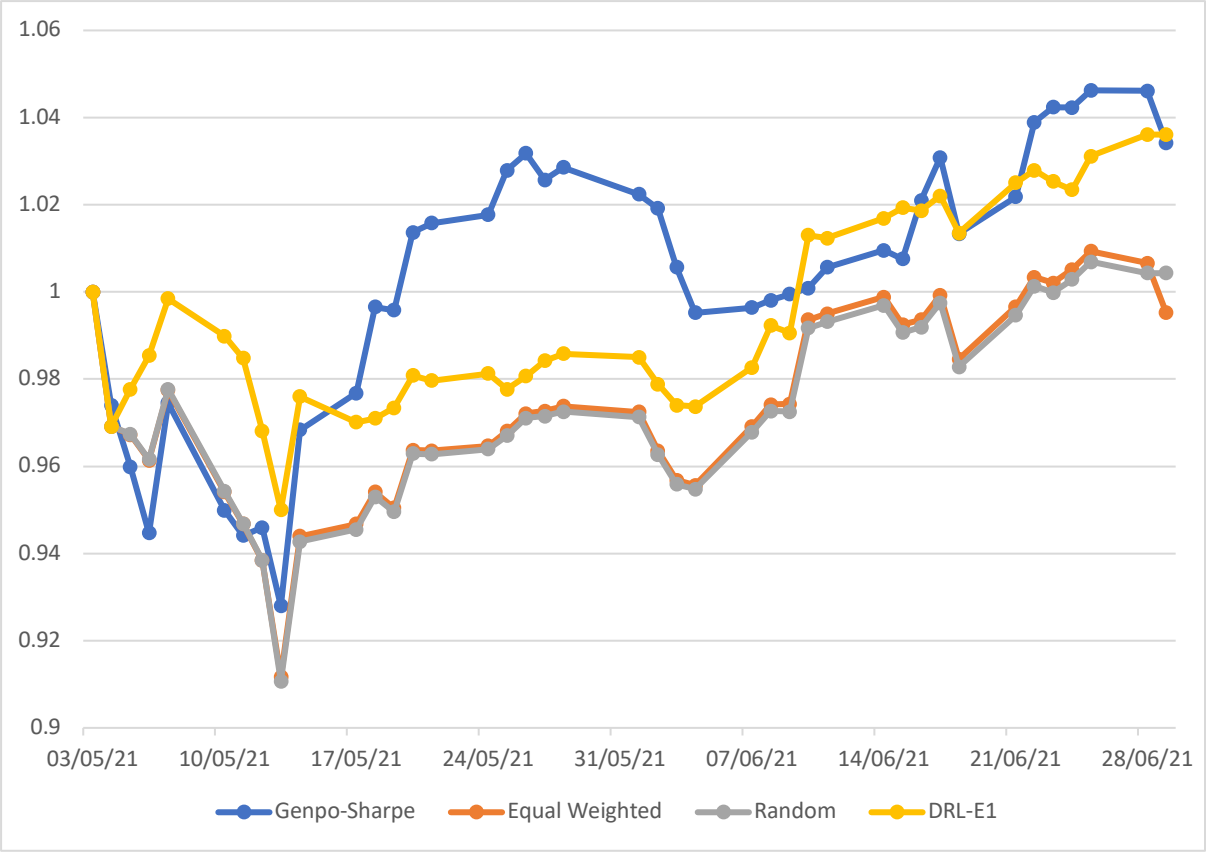

Figura 17 Rendimiento de estrategias para un portafolio de activos del sector Salud de NYSE en el periodo del 1 mayo de 2021 al 30 de junio de 2021

En la figura 17 podemos observar un rendimiento del capital acumulado de 5%, del agente de DRL con la versión del modelo del entorno E1 propuesto en esta tesis, por encima de la estrategia GenPo-Sharpe. Sin bien el valor final del capital invertido con la estrategia DRL-E1 no difiere porcentualmente de manera considerable con respecto a su mejor competidor, dicha estrategia es preferible en términos de Ratio Sharpe (rendimiento en relación riesgo de la estrategia) con un 38% de diferencia positiva, además de contar con un drawdown máximo 31% inferior.

Este experimento consiste en la comercialización de diez activos del sector Energía de NYSE. A continuación, los resultados:

	Retorno anualizado %	**Retorno acumulado %**	**Volatilidad anualizada %**	**Ratio de Sharpe**	**Drawdown máximo %**
GenPo-Sharpe	63.45	8.11	14.97	3.36	-3.44
Equal Weighted	130.34	14.16	18.97	4.50	-3.79
Random	3.86	0.60	19.20	0.29	-8.95
E1	84.90	10.25	20.95	3.04	-4.23
E2	173.38	17.31	21.84	4.72	4.45
E3	116.41	13.04	17.32	4.55	3.21
E4	146.33	15.38	28.48	3.31	-6.22
E5	109.84	12.48	17.94	4.22	-4.55
E6	81.07	9.88	18.03	3.38	-3.65
E7	107.75	12.31	21.38	3.53	-4.64
E8	124.33	13.68	23.42	3.57	-5.43

Tabla 9 Resultados de estrategias de trading para un portafolio de activos pertenecientes al sector Energía de NYSE

En la tabla 9 y resaltado en color amarillo podemos apreciar los mejores resultados globales de acuerdo a cada métrica evaluada. Siendo, para este caso, la estrategia del agente de DRL con variante del modelo del entorno E2, propuesta como método de esta tesis, la ganadora en las métricas de retorno y dando un rendimiento al final del periodo de 17.31% con respecto al capital inicial.

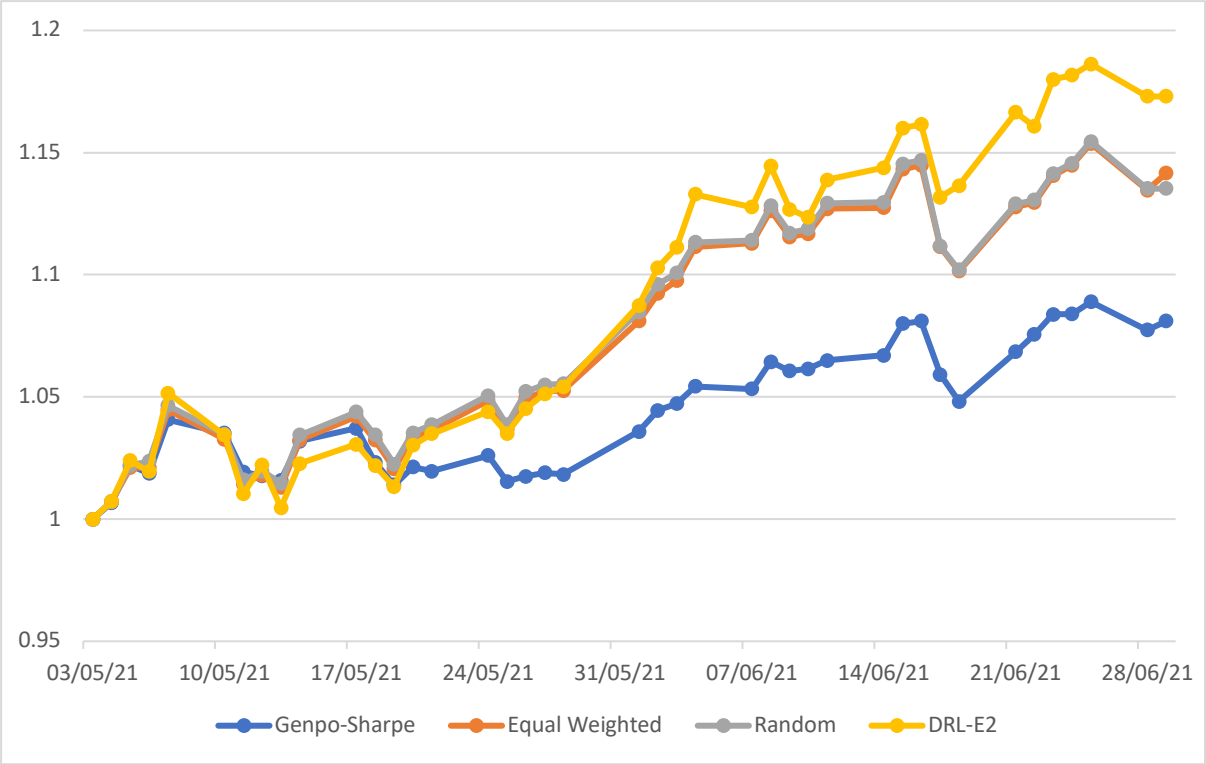

Figura 18 Rendimiento de estrategias para un portafolio de activos del sector Energía de NYSE en el periodo del 1 mayo de 2021 al 30 de junio de 2021

En la figura 18 es posible observar un rendimiento del capital acumulado de 22%, del agente de DRL con la versión del modelo del entorno E2 propuesto en esta tesis, por encima de la estrategia Equal Weighted.

5.2 Agente DRL

5.2.1 Descripción de los datos y diseño del experimento

El agente de DRL, que forma parte del método propuesto en este trabajo de tesis, fue empleado en esta experimentación para la asignación continua de recursos en portafolios de activos de la BSE previamente expuestos en el trabajo de (Conegundes & Pereira, 2020).

De acuerdo con (Conegundes & Pereira, 2020), los portafolios de inversión anuales, utilizados para trading, fueron definidos bajo las siguientes condiciones: primeros 10 activos con mayor peso en el fondo de inversión cotizado (ETF) BOVA11 en la primera sesión de comercialización de los años 2017 a 2019.

2017	2018	2019
ITUB4	ITUB4	ITUB4
BBDC4	VALE3	VALE3
ABEV3	BBDC4	BBDC4
PETR4	ABEV3	PETR4
VALE3	PETR4	ABEV3
BRFS3	B3SA3	BBAS3
BBAS3	ITSA4	B3SA3
ITSA4	BBAS3	ITSA4
B3SA3	UGPA3	LREN3
UGPA3	BRFS3	UGPA3

Tabla 10 Los 10 mejores activos de BOVA11

Los datos que alimentan al agente de DRL consisten de series de tiempo de dos años de antigüedad, con respecto al año de selección del portafolio, de los precios de cierre de cada uno de los activos que componen al portafolio.

La configuración del agente y entorno para cada año se compuso de la siguiente forma:

Año	Entorno	Agente
2017	**Variables:** Open, Close, Volume **Indicadores técnicos:** sma_[20], rsi_14, boll_20 **Función de recompensa:** PV	• 100 neuronas capas internas de MLPs • Funciones de activación ReLU
2018		• 75 neuronas capas internas de MLPs • Funciones de activación ReLU • Tanh como función de activación para la capa de salida del critic
2019		• 90 neuronas capas internas de MLPs • Funciones de activación ReLU

Tabla 11 Variantes de modelo del mercado para agente de DRL y configuración

5.2.2 Resultados

En las siguientes tablas se exponen los resultados de agente de DRL en comparación con las tres configuraciones del método propuesto en (Conegundes & Pereira, 2020) y el índice Ibovespa:

	Retorno anualizado %	**Retorno acumulado %**	**Volatilidad anualizada %**	**Ratio de Sharpe**	**Drawdown máximo %**
C-DRL2	32.03	31.3	25.3	1.22	-22.84
C-DRL3	28.21	26.2	23.63	1.17	-27.35
C-DRL5	38.16	37.27	20.47	1.68	-18.73
Ibovespa	26.59	26.9	13.1	1.87	-12
Agente DRL	46.40	46.63	19.72	2.03	-9.62

Tabla 12 Rendimiento de estrategias para portafolio BOVA11-2017 en el periodo del 1 enero de 2017 al 31 de diciembre de 2017

En la tabla 12 y resaltado en color amarillo podemos apreciar los mejores resultados globales de acuerdo a cada métrica evaluada. Siendo, para este caso, la estrategia del agente de DRL, propuesto en este trabajo de tesis, la ganadora en las métricas de retorno y dando un rendimiento al final del periodo de 46.65% con respecto al capital inicial, un 74% por encima de Ibovespa.

	Retorno anualizado %	**Retorno acumulado %**	**Volatilidad anualizada %**	**Ratio de Sharpe**	**Drawdown máximo %**
C-DRL2	63.39	61.8	25.9	2.03	-18.63
C-DRL3	39.81	38.88	22.04	1.63	-18.16
C-DRL5	120.7	117.26	28.24	2.95	-17.52
Ibovespa	15.28	14.96	22.33	0.75	-20.4
Agente DRL	22.08	21.31	27.53	0.86	-23.68

Tabla 13 Rendimiento de estrategias para portafolio BOVA11-2018 en el periodo del 1 enero de 2018 al 31 de diciembre de 2018

En la tabla 13 se puede observar que la mejor estrategia se derivó el agente C-DRL5 propuesto por (Conegundes & Pereira, 2020) con un retorno acumulado de 117.26% con respecto al capital de inicio. Si bien nuestro agente de DRL se encuentra lejano a este resultado, en comparación con el índice Ibovespa nos encontramos 42% por encima.

	Retorno anualizado %	Retorno acumulado %	Volatilidad anualizada %	Ratio de Sharpe	Drawdown máximo %
C-DRL2	98.84	92.28	24.3	2.87	-11.12
C-DRL3	51.44	50.2	19.16	2.26	-15.48
C-DRL5	8.41	8.23	20.27	0.5	-29.82
Ibovespa	34.29	31.8	13.38	2.27	-10
Agente DRL	38.75	37.67	21.10	1.66	-9.17

Tabla 14 Rendimiento de estrategias para portafolio BOVA11-2019 en el periodo del 1 enero de 2019 al 31 de diciembre de 2019

Para finalizar, en la tabla 14 una de las propuestas (C-DRL2) de (Conegundes & Pereira, 2020) obtiene el mayor retorno acumulado con un porcentaje de ganancia del 92.28% con respecto al capital de inicio. El agente de DRL presentado en este trabajo se comporto ligeramente mejor que el índice Ibovespa con un 18% de diferencia positiva en términos del retorno acumulado.

6 Conclusiones y trabajos futuros

En este trabajo fue propuesto el método para trading algorítmico GA-DRL, el cual consiste de dos etapas principales: (1) el algoritmo genético Genpo-Sharpe para la preselección de un portafolio de inversión y (2) un agente de aprendizaje por refuerzo profundo para el trading diario del portafolio obtenido en la primera etapa.

Derivado de la experimentación realizada con activos de los sectores Tecnología, Salud y Energía de la NYSE, se observo, en los tres sectores, una ventaja real del incremento del valor inicial del capital (retorno acumulado) del 195%, 5% y 22% respectivamente contra el mejor resultado de un par de estrategias Buy & Hold (Genpo-Sharpe y Equal Weighted) y una estrategia totalmente aleatoria. Teniendo además como refuerzo para la elección de las estrategias derivadas del método DRL-GA presentado en este trabajo las métricas Ratio de Sharpe y drawdown máximo.

Una experimentación complementaria con portafolios derivados del ETF BOVA11 de la BSE mostraron un comportamiento positivo, en comparación con el índice Ibovespa, del agente de DRL también propuesto en este trabajo, siendo la ventaja relativa de 74%, 44% y 18% en los años 2017, 2018 y 2019 respectivamente.

El método GA-DRL expuesto en esta tesis se presenta como una opción más, con resultados prometedores, para la asignación continua de recursos financieros o trading financiero en la NYSE y BSE.

6.1 Trabajos futuros

A continuación, se enumeran de manera no exhaustiva posibles trabajos futuros derivados de esta tesis:

1. Diseño y desarrollo de un agente de aprendizaje por refuerzo para trading algorítmico con diferentes arquitecturas de redes neuronales tales como CNN, LSTM, etc.
2. Diseño y desarrollo de un agente de aprendizaje por refuerzo para trading algorítmico en intervalos de actuación muy cortos (horas, minutos, segundos).
3. Implementación de modelos de costo de transacción en un agente de aprendizaje por refuerzo para trading algorítmico.
4. Diseño y desarrollo de algoritmos para ejecución de señales de trading emitidas por un agente de aprendizaje por refuerzo.
5. Implementación de algoritmos de ajuste de híper parámetros para el método GA-DRL.

Referencias

Darškuvienė, V. (2010). *Financial Markets.* Vytautas Magnus University.

Meyer, G. (23 de abril de 2019). *Automation is the future of futures markets* . Obtenido de Financial Times: https://www.ft.com/content/4d589796-6211-11e9-a27a-fdd51850994c

Treleaven, P., Galas, M., & Lalchand, V. (2013). Algorithmic trading review. *Commun. ACM*, 76–85.

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

Branco, J. (2017). *Reinforcement Learning Applied to Forex Trading.* Técnico Lisboa.

Dybvig, P. H., & Ross, S. A. (2003). *Arbitrage, State Prices and Portfolio Theory Handbook of the Economics of Finance.*

Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does Algorithmic Trading Improve Liquidity? *The Journal Of Finance*.

Sharpe, W. F. (1994). The Sharpe Ratio. *The Journal of Porfolio Managment*, 49-58.

Li, Y., Zheng, W., & Zheng, Z. (2019). Deep Robust Reinforcement Learning for Practical Algorithmic Trading. *IEEE Access*.

Ponomarev, E. S., Oseledets, I. V., & Cichocki, A. S. (2019). Using Reinforcement Learning in the Algorithmic Trading Problem. *Journal of Communications Technology and Electronics*, 1450-1457.

Conegundes, L., & Pereira, A. C. (2020). Beating the Stock Market with a Deep Reinforcement Learning Day Trading System. *International Joint Conference on Neural Networks (IJCNN)* , 1-8.

Lei, K., Zhang, B., Li, Y., Yang, M., & Shen, Y. (2020). Time-driven feature-aware jointly deep reinforcement learning for financial signal representation and algorithmic trading. *Expert Systems with Applications*.

Yuan, Y., Wen, W., & Yang, J. (2020). Using Data Augmentation Based Reinforcement Learning for Daily Stock Trading. *Electronics*.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, 279-292.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 436-444.

Rosenblatt, F. (1960). Perceptron Simulation Experiments . *Proceedings of the IRE*, 301–309.

Zai, A., & B., B. (2020). *Deep Reinforcement Learning in Action.* Manning Publications.

Statista Research Department. (1 de Febrero de 2021). *Largest stock exchange operators worldwide as of February 2021, by market capitalization of listed companies*. Obtenido de Statista: https://www.statista.com/statistics/270126/largest-stock-exchange-operators-by-market-capitalization-of-listed-companies/

Frausto, J., Gonzalez, J., Castilla, G., Purata, J., Soto, D., & Hernandez, L. (2020). GenPo-Sharpe: Stock Selection for Investing Portfolio usinga Genetic Algorithm with Sharpe Ratio Applied to Mexican Stock Exchange. *NEO*, (págs. 92-93).

Dash, R., & Dash, P. K. (2016). A hybrid stock trading framework integrating technical analysis with machine learning techniques. *The Journal of Finance and Data Science*, 42–57.

W., B., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE*.

Vanstone, B., & Finnie, G. (2009). n empirical methodology for developing stockmarket trading systems using artificial neural networks. *Expert Systems with Applications*, 6668–6680.

Schulman, J. W. (2017). Proximal Policy Optimization Algorithms. *Computing Research Repository* .

Mnih, V., Puigdomenech, A., Mirza, M., Graves, A., Lillicrap, T., Harley, T., . . . Kavukcuoglu, K. (2016). Asynchronous Methods for Deep Reinforcement Learning. *Proceedings of The 33rd International Conference on Machine Learning*, 1928-1937.