RACEF
Revista de Administração, Contabilidade e Economia da **Fundace**

FUNDACE

Ribeirão Preto, Novembro de 2024
Edição: v. 15, n.3 (2024)

ALOCAÇÃO DE FUNDOS DE INVESTIMENTO IMOBILIÁRIOS E ESTRATÉGIAS DE NEGOCIAÇÃO: CARTEIRAS ELABORADAS EM ALGORITMOS DE *REINFORCEMENT LEARNING*

REAL ESTATE INVESTMENT FUND ALLOCATION AND TRADING STRATEGIES: PORTFOLIOS DEVELOPED USING REINFORCEMENT LEARNING ALGORITHMS

DOI: HTTP://DX.DOI.ORG/10.13059/RACEF.V15I3.1216

Julia Pinheiro Barboza
juliapinheirobarboza@gmail.com
Universidade Federal de Uberlândia

Gustavo Carvalho Santos
gustavocavsantos@gmail.com
Universidade Federal de Uberlândia

Daniel Vitor Tartari Garruti
garrutidaniel@gmail.com
Universidade Federal de Uberlândia

Flávio Barboza
flmbarboza@ufu.br
Universidade Federal de Uberlândia

Data de envio do artigo: 19 de Janeiro de 2024.
Data de aceite: 04 de Novembro de 2024.

Resumo: Este estudo investiga estratégias de negociação em Fundos de Investimento Imobiliários (FIIs) utilizando Reinforcement Learning (RL) para maximizar o retorno de uma carteira. Foram implementados cinco algoritmos de RL, configurados com o método ator-crítico, para gerar estratégias de negociação, que foram avaliadas e comparadas com métodos tradicionais, como Buy and Hold (B&H), mínima variância e o Índice IFIX. A amostra final, composta por 26 fundos, foi selecionada considerando critérios como histórico mínimo de 5 anos e volume de negociação. As estratégias de RL superaram a mínima variância e tiveram desempenho comparável ao B&H, mas não superaram o IFIX. Isso sugere que o mercado de FIIs pode ser eficiente, limitando a eficácia de técnicas avançadas como RL e dificultando a superação do índice. As contribuições do estudo incluem demonstrar o potencial do RL em superar estratégias tradicionais em determinados contextos, oferecendo possíveis compreensões para investidores, gestores de carteiras e literatura científica.

Palavras-chave: Ator Crítico; Fundos de Investimento Imobiliários; Estratégias de Negociação; Aprendizado por Reforço; Gestão de Portfólio.

Abstract: *FThis study investigates trading strategies in Real Estate Investment Funds (REITs) using Reinforcement Learning (RL) to optimize portfolio returns. Five RL algorithms, configured with the actor-critic method, were implemented to generate trading strategies, which were evaluated against traditional methods such as Buy and Hold (B&H), minimum variance, and the IFIX Index. The final sample consisted of 26 funds, selected based on criteria including a minimum historical period of five years and trading volume. The results indicate that RL strategies outperformed the minimum variance approach and showed performance comparable to B&H, but did not exceed the IFIX Index. This finding suggests that the REIT market may be efficient, limiting the effectiveness of advanced techniques like RL. The study contributes to demonstrating RL's potential to surpass traditional strategies in specific contexts, providing possible understandings for investors, portfolio managers, and the academic literature.*

Keywords*: Actor-Critic; Real Estate Investment Trusts; Trading Strategies; Reinforcement Learning; Portfolio Management.*

1 INTRODUÇÃO

Markowitz (1952) desenvolveu a Teoria Moderna de Portfólio, defendendo a diversificação de investimentos como forma de reduzir riscos e aumentar retornos. Nesse cenário, Fama (1970) apresentou a Hipótese de Mercado Eficiente (HME), que afirma que os preços dos ativos refletem todas as informações disponíveis, o que impede a obtenção de lucros superiores ao índice de mercado. Em contraponto, Lo (2004) introduziu a Hipótese de Mercado Adaptativo (HMA), argumentando que os investidores apresentam vieses e que, em momentos de incerteza ou rápidas mudanças, é possível prever preços futuros e obter retornos consideráveis. Ambas as teorias serão exploradas e relacionadas à gestão de portfólio proposta neste estudo, possibilitando um debate sobre qual delas sustenta de maneira mais coerente a estratégia adotada.

Os Fundos de Investimento Imobiliário (FII) surgiram nos Estados Unidos na década de 1880, ganhando relevância na década de 1960 com o Real Estate Investment Trust (REIT) (Castello Branco; Monteiro, 2003). A legislação equiparou seus benefícios fiscais aos dos mutual funds, resultando em transformações regulatórias significativas, como a Lei de Reforma Tributária de 1986, a Lei de Orçamento e Reconciliação Omnibus de 1993, e a Lei de Modernização do REIT de 1999, impactando a indústria em tamanho e composição (Feng et al., 2011; Geltner et al., 2001; Chan et al., 2003). Os REITs cresceram de US$ 26 bilhões em 1993 para mais de US$ 400 bilhões em 2006 (Feng et al., 2011).

No Brasil, os Fundos Imobiliários começaram a ganhar impulso na década de 1990, com a legalização em junho de 1993 pela

Lei 8.668, que estabeleceu a estrutura e o regime tributário, ajudando a desenvolver os setores imobiliário e agroindustrial. A Instrução 205 da CVM (1994), foi crucial na regulamentação desses fundos, garantindo transparência e segurança aos investidores. Em 2018, o mercado de fundos imobiliários no Brasil contava com 368 fundos, totalizando R$ 80,80 bilhões em patrimônio líquido e R$ 45 bilhões em valor de mercado. Em 2023, esse número aumentou para 819 fundos, com patrimônio líquido superior a R$ 200 bilhões e valor de mercado de R$ 138 milhões, evidenciando o crescimento e potencial do mercado. Essa ascensão motivou a escolha dos FIIs como foco deste estudo.

Santos et al. (2023) investigaram a otimização de carteiras de investimento com algoritmos de aprendizado por reforço (RL), analisando 40 ações e 10 commodities em mercados globais. O estudo destacou que a abordagem proposta superou benchmarks e algoritmos populares, gerando retornos mais altos com menor risco, inclusive durante a crise da COVID-19. A pesquisa ressaltou a importância de incluir custos de transação no treinamento do modelo, indicando que a adição de derivativos de commodities pode melhorar significativamente o desempenho da carteira, reduzindo a volatilidade e oferecendo oportunidades atraentes para investidores.

Este estudo apresenta os resultados de desempenho de cinco carteiras elaboradas, evidenciando o desempenho de carteiras construídas com algoritmos de RL em comparação a estratégias tradicionais, e avaliando o impacto de eventos econômicos e políticos. A pesquisa oferece uma contribuição significativa para gestores e investidores que buscam aprimorar suas estratégias para maximizar lucros. Apesar de promissores, os resultados das estratégias de RL em comparação às tradicionais não superaram o desempenho do IFIX, refletindo a complexidade e eficiência do mercado de FIIs, e sugerindo áreas para futuras pesquisas.

Este artigo é estruturado em cinco seções. Após esta introdução, é apresentado o referencial teórico, abrangendo a Teoria de Carteiras e o uso de *Reinforcement Learning* na gestão de carteiras. A seção seguinte detalha os procedimentos metodológicos, incluindo a seleção dos fundos imobiliários e a descrição dos algoritmos utilizados. Na seção 4, os resultados obtidos com os diferentes algoritmos são discutidos, e a seção 5 apresenta a conclusão do trabalho.

2 REFERENCIAL TEÓRICO

2.1 Teoria de Carteiras

A teoria de carteiras, formulada por Harry Markowitz em 1952, destaca-se como fundamental no campo financeiro. Segundo Markowitz (1952), a teoria postula que investidores, buscando maximizar a utilidade esperada de seu patrimônio, consideram tanto o retorno quanto o risco dos ativos em suas decisões. Com a premissa de que os investidores são avessos ao risco, a teoria permite avaliar o desempenho da carteira em termos de retorno e risco, estabelecendo uma relação positiva entre ambos. Dessa forma, os investidores têm a flexibilidade de escolher carteiras que minimizem o risco para um determinado nível de retorno ou maximizem o retorno para um dado nível de risco.

Dessa forma, a diversificação, essencial na teoria de carteiras, é uma estratégia fundamental para reduzir o risco total da carteira, permitindo ao investidor combinar ativos com baixa correlação. Conforme Markowitz (1952), a diversificação é viável devido à correlação negativa entre ativos. Ele propôs a mensuração do risco como a variância dos retornos, uma medida que captura a dispersão em torno da média, possibilitando a comparação de ativos ou carteiras em termos de risco. Essa abordagem é crucial para a avaliação de risco e desempenho das carteiras.

A relação risco-retorno é crucial para verificar se o retorno obtido compensa o risco assumido. No contexto dos FIIs, a aplicação da teoria de carteiras possibilita a seleção de uma carteira ótima desses ativos. O modelo de Markowitz (1952) é aplicável a diversos tipos de ativos, incluindo os imobiliários, desde que

se tenha conhecimento do retorno esperado, variância dos retornos e covariância entre os ativos. Dessa forma, o modelo auxilia na gestão de carteiras de fundos imobiliários, permitindo ao investidor diversificar seus investimentos e reduzir o risco total da carteira. Adicionalmente, medidas de risco como o Coeficiente Beta e o Índice de Sharpe (1964) são empregadas para avaliar o desempenho ajustado ao risco de um portfólio.

Assaf Neto (2019) destaca a relação direta entre risco e retorno nos Fundos de Investimento, onde a busca por maiores rendimentos implica em maior exposição ao risco para o investidor. Contrariamente, fundos que oferecem maior segurança geralmente apresentam retornos mais moderados. A escolha da relação risco-retorno é uma decisão individual do investidor, guiada por sua aversão ao risco. Os principais tipos de risco nos Fundos de Investimentos, como o risco de crédito, risco de mercado, risco de liquidez e risco sistêmico, compõem o cenário desse contexto financeiro.

2.2 Hipóteses do Mercado Financeiro

A HME, formulada por Fama (1970), parte do princípio de que os preços dos ativos financeiros incorporam todas as informações disponíveis. Esta hipótese é categorizada em três formas distintas: fraca, semi-forte e forte. A forma fraca postula que o mercado reflete todas as informações públicas disponíveis. A forma semi-forte inclui a forma fraca e sugere que novas informações são incorporadas instantaneamente pelo mercado. Já a forma forte abarca as duas anteriores, afirmando que os preços refletem todos os tipos de informações, tanto públicas quanto privadas.

Lo (2004) desenvolveu a HMA, que se fundamenta na análise de mercados e pode ser interpretada como uma evolução da HME de Fama (1970). A HMA incorpora princípios evolucionários à economia, baseando-se em leis biológicas como seleção natural, adaptação, mutação e aprendizado para guiar as estratégias e heurísticas de tomada de decisão mais apropriadas. Essa hipótese reconcilia a estrutura neoclássica da HME com o comportamento não ótimo do agente, considerando novas abordagens na tomada de decisão financeira, como aprendizado, adaptação e vieses comportamentais (Burnham, 2013). Consequentemente, segundo a HMA, é possível antecipar os preços dos ativos no mercado financeiro.

2.3 Trabalhos relacionados

Scolese et al. (2015) investigaram o retorno dos FIIs no Brasil entre 2011 e 2015, analisando seu comportamento em relação a índices de renda fixa, variável e imobiliário. Os resultados indicaram maior correlação com juros prefixados e o mercado imobiliário. Feng et al. (2011) estudaram os REITs de capital aberto, abordando o crescimento da indústria, mudanças no foco de propriedade, aumento da alavancagem e flutuações no fluxo de caixa. Iório et al. (2015) compararam o desempenho de três portfólios de FIIs entre 2011 e 2013: um portfólio baseado na teoria de Markowitz (1952), outro com pesos iguais e o IFIX. Apesar de retornos semelhantes, o IFIX apresentou melhor desempenho no balanço risco-retorno.

Yang et al. (2020) propuseram uma estratégia integrada de negociação automatizada usando aprendizado por reforço profundo, combinando PPO, A2C e DDPG, superando as versões individuais dos algoritmos, especialmente em termos de retorno ajustado ao risco. Santos et al. (2023) exploraram a otimização de carteiras com RL, incluindo ações de mercados globais e commodities, mostrando que a abordagem superou benchmarks tradicionais, com melhores retornos e menor risco, inclusive ao considerar custos de transação. Sun et al. (2023) apresentaram uma visão abrangente do uso de aprendizado por reforço em negociação quantitativa, destacando modelos da biblioteca FinRL para identificar oportunidades de investimento.

2.4 *Reinforcement Learning* para gestão de carteiras

O *Reinforcement Learning* (RL), um subcampo do machine learning, treina um agente para maximizar recompensas cumulativas, sendo aplicado em negociações (Yang et al., 2020). Yang et al. (2020) descrevem uma estratégia automatizada de negociação usando aprendizado profundo por reforço, integrando técnicas de RL para criar um modelo robusto diante das incertezas do mercado. A utilização de múltiplos agentes de RL aprimora a capacidade de adaptação a diferentes condições, com resultados experimentais que comprovam sua eficácia.

Sun et al. (2023) oferecem uma visão abrangente dos métodos de RL aplicados à negociação quantitativa, destacando desafios e oportunidades. O estudo demonstra o potencial do RL para desenvolver estratégias adaptativas que se ajustam às mudanças de mercado.

3 PROCEDIMENTOS METODOLÓGICOS

Os dados relativos a cada um dos FIIs foram coletados de janeiro de 2018 a maio de 2023, sendo divididos aproximadamente em 70% para treinamento e 30% para teste, sendo aplicada uma estratificação temporal, ou seja, a amostra treino abarca dados entre janeiro de 2018 a dezembro de 2021 e a amostra teste é composta por dados mais recentes, de janeiro de 2022 até maio de 2023. Esse critério de divisão é amplamente utilizado em estudos com séries temporais para garantir que os dados de teste reflitam informações não vistas durante o treinamento.

A proporção de 70-30 é comum na literatura, como sugerido Zhang et al. (2020), embora Sun et al. (2023) mencione que esse ponto é ainda considerado abrangente, tendo estudos que aplicam 50-50, como Santos et al. (2023) até 90-10. Entretanto, a divisão aplicada tende a evitar a ocorrência de sobreajuste. As informações coletadas incluem os preços de abertura, fechamento, máximos e mínimos diários dos fundos imobiliários, volume de negociação, juntamente com os indicadores da análise técnica, conforme descrito por Santos et al. (2023). Adicionalmente, foram obtidos dados relacionados aos segmentos dos FIIs.

Cabe mencionar que foram aplicadas as configurações padrão recomendadas pela biblioteca FinRL para todos hiperparâmetros de cada um dos algoritmos (A2C, PPO, DDPG, SAC e TD3). Essa escolha foi baseada na robustez das configurações iniciais para problemas de negociação financeira, conforme descrito por Liu et al. (2020).

3.1 Construção da amostra do estudo

De acordo com informações da [B]³, o IFIX é construído a partir de uma carteira teórica de ativos, seguindo critérios definidos no Manual de Definições e Procedimentos dos Índices da B3 (2023). Classificado como um índice de retorno total, o IFIX visa ser um indicador do desempenho médio das cotações dos fundos imobiliários na B3. A seleção dos fundos, orientada por critérios específicos, resultou na amostra inicial de 111 fundos.

A exigência de um histórico mínimo de 5 anos para os fundos imobiliários está relacionada à aplicação do método de RL, pois uma análise histórica abrangente permite ao algoritmo identificar padrões de desempenho ao longo do tempo. Utilizando um período mais extenso, o RL pode incorporar mais informações passadas, aumentando a precisão e confiabilidade da análise e reduzindo a influência de flutuações de curto prazo.

Além disso, a exclusão dos FOFs (Fundos de Fundos) se deu pela disponibilidade de dados de fundos que não incluíssem outros em seus portfólios, visto que o estudo analisa uma carteira com características específicas. Isso resultou na redução da amostra para 40 fundos elegíveis para a pesquisa.

A seleção de ativos líquidos para o IFIX está ligada à preferência por ativos de fácil negociabilidade, pois a liquidez é crucial para a análise e acompanhamento desses fundos, favorecendo uma formação de preços eficiente.

Essa característica é vital para o sucesso do algoritmo de RF, que aprende padrões a partir de altos volumes de negociação (Oshingbesan et al., 2022; Zhang et al., 2019). A abundância de dados dos ativos líquidos auxilia na tomada de decisões (Hambly et al., 2023). Entre os fundos elegíveis, 26 apresentam alto volume de negociações e foram selecionados para a amostra final, conforme mostrado no Quadro 1.

Quadro 1 - FIIs selecionados para a amostra da pesquisa

Código	NOME DO ATIVO	TIPO ANBIMA	SEGMENTO ANBIMA
ALZR11	Alianza Trust Renda Imobiliária	Híbrido (Tipo: Renda)	Misto
BBPO11	BB Progressivo II	Lajes Corporativas (Tipo: Renda)	Agências de Bancos
BCRI11	Banestes Recebíveis Imobiliários	Títulos e Valores Mobiliários	Papel
BRCR11	BTG Pactual Corporate Office	Híbrido	Lajes Corporativas
BTLG11	BTG Pactual Logística FDO INV IMOB – FII	Híbrido (Tipo: Renda)	Imóveis Industriais e Logísticos
CARE11	Brazilian Graveyard and Death Care	Híbrido (Tipo: Títulos e Valores Mobiliários)	Misto
CPTS11	Capitania Securities II	Títulos e Valores Mobiliários	Papel
FIIB11	Industrial do Brasil	Híbrido	Imóveis Industriais e Logísticos
GGRC11	GGR Covepi Renda	Logística (Tipo: Híbrido)	Imóveis Industriais e Logísticos
HGBS11	CSHG Brasil Shopping	Shoppings (Tipo: Renda)	Shoppings
HGCR11	CGHG Recebíveis Imobiliários	Títulos e Valores Mobiliários	Papel
HGLG11	CGHG Logística	Logística (Tipo: Renda)	Imóveis Industriais e Logísticos
HGRE11	CSHG Real Estate	Lajes Corporativas (Tipo: Renda)	Lajes Corporativas
JSRE11	JS Real Estate Multigestão	Híbrido	Misto
KNCR11	Kinea Rendimentos Imobiliários	Títulos e Valores Mobiliários	Papel
KNIP11	Kinea Índice de Preços	Títulos e Valores Mobiliários	Papel
KNRI11	Kinea Renda Imobiliária	Híbrido (Tipo: Renda)	Misto
MFII11	Mérito Desenvolvimento Imobiliário	Híbrido	Fundo de Desenvolvimento
MXRF11	Maxi Renda	Híbrido	Papel
NSLU11	Hospital Nossa Sra Lourdes	Hospital (Tipo: Renda)	Hospitalar
OUJP11	Ourinvest JPP	Híbrido (Tipo: Títulos e Valores Mobiliários)	Papel
PORD11	Polo Recebíveis Imobiliários II	Títulos e Valores Mobiliários	Papel
SDIL11	SDI Logística Rio	Logística (Tipo: Renda)	Imóveis Industriais e Logísticos
SPTW11	SP Downtown	Lajes Corporativas (Tipo: Renda)	Lajes Corporativas
VISC11	Vinci Shopping Centers	Shoppings (Tipo: Renda)	Shoppings
VRTA11	Fator Verita	Títulos e Valores Mobiliários	Papel

Fonte: Dados da pesquisa.

3.2 Coeficientes de análise

Proposto por Sharpe (1963), o modelo de análise de carteiras baseia-se na premissa de que investidores avessos ao risco buscam maximizar a utilidade esperada de seus retornos. O coeficiente beta é a razão entre a covariância do retorno da carteira e o retorno do mercado, dividida pela variância deste. A pesquisa de Sharpe mostra que, em equilíbrio, o retorno esperado de uma carteira é linearmente relacionado ao seu beta.

Segundo Assaf Neto et al. (2008), a carteira de mercado, com beta de 1,0, é a mais diversificada e reflete apenas o risco sistemático. Ativos com beta de 1,0 têm retorno igual à média do mercado; acima de 1,0, indicam maior risco e expectativa de retorno superior; abaixo de 1,0, sinalizam menor risco e retorno inferior.

O Índice de Sharpe (1964) avalia o desempenho de um investimento em relação ao seu risco, quantificando a relação entre o retorno excedente do ativo e sua volatilidade. Um Índice de Sharpe maior indica um retorno ajustado ao risco mais atrativo e pode ajudar a analisar desempenhos passados e prever futuros.

A volatilidade, medida de risco que reflete as oscilações dos preços dos ativos em torno da média (Markowitz, 1952), indica maior risco quando mais alta. Essa abordagem permite ao investidor escolher uma carteira que minimize o risco para um nível de retorno ou maximize o retorno para um nível de risco. A volatilidade pode ser afetada por fatores econômicos, políticos e sociais da região dos imóveis.

O risco é crucial na seleção e gestão de carteiras de FIIs ou REITs. Fatores como a qualidade dos imóveis, a solidez do inquilino e a competência do gestor influenciam esse aspecto. Liow et al. (2019a) ressaltam a importância da volatilidade nas decisões de investimento em imóveis comerciais, enquanto Liow et al. (2019b) investigam a relação entre a volatilidade dos preços de REITs em diferentes países.

3.3 Algoritmos utilizados

Na presente pesquisa, os principais objetos de análise foram os cinco algoritmos de RL: Advantage Actor-Critic (A2C), Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradient (DDPG), Soft Actor-Critic (SAC) e Twin Delayed Deep Deterministic Policy Gradient (TD3). Esses métodos foram selecionados por representarem padrões disponíveis na coleção FinRL, um arcabouço desenvolvido com o propósito de estudo e aplicação no setor financeiro, além de se basear nas pesquisas de Yang et al. (2020); Santos et al. (2023) e Sun et al. (2023).

O método A2C, que combina as funções de ator e crítico, utiliza uma função de vantagem para estabilizar o gradiente de políticas, melhorando o processo de aprendizado (Mnih et al., 2016). Ao integrar política e valor, o A2C consegue aprender uma política ideal e se adaptar a espaços de ação dinâmicos, o que é útil em ambientes com múltiplos agentes (Wang et al., 2016).

Entre suas vantagens, destaca-se a estabilidade, pois a combinação de métodos baseados em políticas e valores contribui para um aprendizado mais estável (Wang et al., 2016). O A2C é eficiente em termos de amostragem, utilizando a função de vantagem para reduzir o viés e demandando menos interações com o ambiente. Além disso, sua simplicidade de implementação o torna uma escolha viável em comparação a métodos mais complexos, como os off-policy. Outra característica importante é sua boa convergência, devido à sincronização das atualizações dos múltiplos atores, evitando a instabilidade de atualizações assíncronas.

Por outro lado, o A2C tem desvantagens. É computacionalmente mais caro, pois tanto o ator quanto o crítico precisam ser atualizados a cada passo, o que aumenta o processamento. Como um método on-policy, pode ser menos eficiente em ambientes grandes e complexos quando comparado a alternativas off-policy, como DDPG ou SAC. Além disso, embora o A2C reduza a variância em relação aos métodos de gradiente de política puro, ainda pode sofrer de

alta variância por conta de sua natureza *on-policy.*

O DDPG, visando maximizar o retorno do investimento (Lillicrap et al., 2015), integra frameworks de *Q-learning* (Learning et al., 1998) e *policy gradient* (Sutton et al., 2000), incorporando redes neurais para mapeamento. Assim, o DDPG extrai padrões diretamente das observações por meio do gradiente político, modelando estados para ações de maneira determinística para se adaptar eficientemente a ambientes de espaços dinâmicos.

Em conformidade com Dulac-Arnold et al. (2020), a cada etapa, o agente DDPG realiza uma ação at em st, recebe uma recompensa rt e chega em st+1. As transições (st, at, st+1, rt) são armazenadas na área de memória temporária de repetição R. Um conjunto de N transições é selecionado de R e o valor-Q yi é atualizado como:

$$yi = ri + \gamma Q'(si{+}1, \mu'(si{+}1 \,|\theta\mu', \theta Q')), i = 1, \cdots, N \qquad (1)$$

Sob a perspectiva de Fang et al. (2019), a rede crítica é atualizada minimizando a função de perda L(θQ), que é a diferença esperada entre as saídas da rede crítica alvo Q' e da rede crítica Q, ou seja,

$$L(\theta Q) = \mathrm{E}[(yi - Q(st, at|\theta Q))\text{^}2] \qquad (2)$$

O DDPG possui vantagens notáveis. Sendo um algoritmo off-policy, ele pode reutilizar amostras de experiências passadas, o que o torna eficiente em ambientes com grandes espaços de ação contínua. Além disso, é um método model-free e determinístico, o que é ideal para tarefas de controle em alta dimensão, como robótica e manipulação de objetos.

Por outro lado, o DDPG também apresenta desafios. A instabilidade no treinamento pode causar oscilações e dificuldades na convergência, especialmente em tarefas complexas. Ele é sensível à escolha de hiperparâmetros, exigindo ajustes cuidadosos para obter bons resultados. O DDPG também é suscetível a ruídos nas ações e observações, o que pode comprometer o desempenho em ambientes com muita variação ou incerteza. Além disso, sendo determinístico, o algoritmo pode não explorar adequadamente o espaço de soluções, o que é uma limitação em problemas que demandam maior exploração (Dulac-Arnold, 2020).

Deste modo, o DDPG é eficiente no tratamento do espaço de ação contínuo e, portanto, é adequado para negociação de Fundos de Investimento Imobiliários (FIIs).

O PPO, de acordo com Schulman et al. (2017), é empregado para gerenciar a atualização do gradiente de política, assegurando que a política atualizada não apresente diferenças notáveis em relação à anterior. O PPO simplifica a abordagem da Trust Region Policy Optimization (TRPO) ao incorporar um termo de corte na função objetivo (Schulman et al., 2015, 2017). Portanto, considerando a proporção de probabilidade entre as políticas antigas e novas expressa por:

$$rt\,(\theta) = \pi\theta\,(at\,|st)\,|\,\pi\theta old\,(at\,|st) \qquad (3)$$

A função objetivo substituta cortada do PPO pode ser entendida pela expressão (Schulman et al., 2017):

$$\mathrm{L}^{\mathrm{CLIP}}(\theta) = \mathrm{E}\hat{}t\,[\min(rt\,(\theta)A\hat{}(st\,, at), \mathrm{clip}(rt\,(\theta), 1 - \epsilon, 1 + \epsilon)A\hat{}(st\,, at))] \qquad (4)$$

Onde rt (θ)Aˆ(st, at) é o objetivo normal do gradiente de política e Aˆ(st, at) é a função de vantagem estimada. A função *clip* (rt (θ), 1− ϵ, 1 + ϵ) corta a proporção rt (θ) para estar dentro de [1− ϵ, 1 + ϵ]. O PPO desestimula alterações significativas de política fora do intervalo cortado. Assim, o PPO aprimora a estabilidade do treinamento das redes de política ao limitar a atualização da política em cada passo de treinamento. Optou-se pelo PPO para negociação de ações por ser estável, veloz e mais fácil de implementar e ajustar.

O PPO possui algumas vantagens, entre elas, está o equilíbrio entre simplicidade e desempenho, com a função objetivo permitindo múltiplas atualizações em minilotes de dados, o que torna o método mais eficiente em termos de amostragem do que abordagens tradicionais de gradiente de política. Além disso, o PPO se diferencia por manter a estabilidade do treinamento através do mecanismo de clipping, que limita grandes mudanças na política, evitando oscilações abruptas no aprendizado. O algoritmo também se destaca em tarefas de controle contínuo e em jogos de Atari, superando outros métodos como A2C e TRPO, sem a complexidade deste último (Schulman et al., 2017).

Por outro lado, o PPO apresenta algumas desvantagens. Embora seja mais simples que o TRPO, ainda requer ajustes precisos de hiperparâmetros, como o valor de ϵ, para garantir que as atualizações sejam eficazes, mas não causem instabilidade. Outro ponto a ser considerado é o risco de overfitting, pois múltiplos passos de otimização nos mesmos dados podem levar a esse problema se não forem bem controlados. Apesar de sua simplicidade em relação ao TRPO, o PPO ainda exige cálculos adicionais, como a função de vantagem, tornando sua implementação mais complexa em comparação a métodos puramente baseados em gradiente de política (Schulman et al., 2017).

O SAC é um algoritmo off-policy de RL que integra a abordagem ator-crítico com otimização da entropia máxima, visando uma exploração mais eficiente e um equilíbrio entre exploração e explotação (Haarnoja et al., 2018). O SAC utiliza a entropia para promover a exploração de informações, adotando uma perspectiva distribucional do objetivo (Ma et al., 2020). O sucesso do SAC em diversas áreas, como robótica e simulações de controle, evidencia sua capacidade em solucionar problemas complexos e de alta dimensão (Han; Sung, 2021). Dessa forma, a eficácia do SAC em diferentes contextos destaca sua aptidão para lidar com os desafios apresentados pelo alto volume de negociações de FIIs.

Conforme Haarnoja et al. (2018), o SAC pode ser calculado por:

$$J(\pi) = E\tau\sim\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(r(st,at) + \alpha H\left(\pi(\cdot\,|st)\right)\right)\right] \qquad (5)$$

onde τ=(s0,a0,s1,a1,...) é uma trajetória gerada pela política π, r(st,at) é a recompensa recebida no tempo t, γ é o fator de desconto, α é um parâmetro que controla o trade-off entre recompensa e entropia, e H(π(°|st)) é a entropia da política π no estado st.

Entre as principais vantagens do SAC, destaca-se sua eficiência na reutilização de dados devido à abordagem *off-policy*, o que acelera o aprendizado. Além disso, o SAC mantém estabilidade durante o treinamento, o que o diferencia de métodos como o DDPG, conhecidos por serem mais instáveis. A maximização da entropia, característica central do SAC, promove políticas estocásticas, garantindo uma exploração eficiente e evitando a convergência prematura a soluções subótimas. Essa robustez o torna adequado para lidar com problemas de controle contínuo em alta dimensão e em ambientes complexos. O SAC também se destaca em benchmarks difíceis, como o controle de humanoides, superando consistentemente outros algoritmos de reforço profundo, tanto off-policy quanto on-policy (Haarnoja et al., 2018).

No entanto, o SAC possui algumas limitações. Ele pode ser sensível à escala de recompensas, exigindo ajustes manuais para manter um bom equilíbrio entre exploração e explotação. Embora apresente alta estabilidade, o algoritmo ainda depende de uma parametrização cuidadosa,

especialmente em relação às funções de valor suavizadas e ao gerenciamento das redes-alvo, para garantir uma convergência estável. Além disso, o SAC pode exigir um custo computacional mais elevado, devido à necessidade de otimizar várias funções simultaneamente, o que pode representar uma desvantagem em termos de tempo de treinamento e recursos necessários (Haarnoja et al., 2018).

O TD3, proposto por Fujimoto et al. (2018), tem como objetivo aprimorar a estabilidade e o desempenho do DDPG (Lillicrap et al., 2015). O TD3 introduz alterações significativas, como a atualização retardada do ator, o uso de duas redes críticas e a incorporação de ruído de ação direcionado. Essas modificações não apenas fortalecem o aprendizado de controle contínuo, mas também contribuem para a estabilidade do processo.

O procedimento inova ao utilizar duas redes fundamentais para estimar a função Q, enfrentando a superestimação comum em algoritmos Q-learning. Ao calcular o valor alvo como o mínimo entre as estimativas das duas redes críticas, o TD3 reduz a propensão à superestimação. Além disso, para aprimorar o desempenho, introduz um atraso na atualização do ator e da rede crítica-alvo, seguindo a proposta de Fujimoto et al. (2018).

A técnica apresenta vantagens, incluindo a redução do viés de superestimação nas estimativas de valor, obtida com o uso de duas redes críticas, o que proporciona uma avaliação mais precisa das ações. A implementação de atualizações de política atrasadas também contribui para a estabilidade do aprendizado, permitindo que as estimativas de valor se consolidem antes de impactar a política. Adicionalmente, a técnica de regularização inspirada no método SARSA resulta em menor variância nas estimativas, elevando ainda mais o desempenho do TD3 em comparação com o DDPG e outros algoritmos de referência em tarefas desafiadoras de controle contínuo (Fujimoto et al., 2018).

Entretanto, o TD3 não está isento de desvantagens. Sua implementação é mais complexa em relação ao DDPG, exigindo a gestão de duas redes críticas e a aplicação de atualizações de política atrasadas, o que pode aumentar a carga computacional e complicar o código. Além disso, a escolha dos hiperparâmetros pode impactar significativamente o desempenho do algoritmo, necessitando de ajustes cuidadosos para otimizar os resultados. Por último, embora o TD3 mostre melhorias em ambientes de controle contínuo, ele pode ter dificuldades em cenários altamente dinâmicos, onde a capacidade de adaptação rápida é essencial (Fujimoto et al., 2018).

4 ANÁLISE DOS RESULTADOS

Os resultados evidenciam um panorama detalhado das estratégias analisadas, com base nas métricas de retorno, retorno acumulado, volatilidade, índice de Sharpe e beta apresentadas na Tabela 1. As estratégias de Reinforcement Learning (RL) alcançaram desempenhos competitivos, especialmente quando comparadas às estratégias tradicionais como Buy and Hold (B&H) e mínima variância. No entanto, é relevante observar que nenhuma das estratégias conseguiu superar o índice de referência IFIX, o que evidencia, mesmo que parcialmente, a Hipótese de Mercado Eficiente proposta por Fama (1970), sugerindo que as informações disponíveis já estão refletidas nos preços dos ativos e que a capacidade de gerar retornos ajustados ao risco superiores ao mercado é limitada.

Complementarmente, este resultado se contrapõe à Hipótese do Mercado Adaptativo (Lo, 2004), que postula que os mercados não são totalmente eficientes, mas sim adaptáveis e sujeitos a períodos de comportamento irracional que podem ser explorados por estratégias dinâmicas, como as de RL. A incapacidade de superar o índice de referência também se alinha aos achados de Iório et al. (2015), que ao analisarem estratégias tradicionais de investimento em fundos imobiliários, não conseguiram superar o IFIX consistentemente. No entanto, vale destacar o contraste com Santos et al. (2023), cujo estudo, embora focado em ações e commodities, identificou oportunidades

de superação do benchmark em determinados setores.

A Tabela 1 revela que, apesar de a estratégia de mínima variância apresentar a menor volatilidade e beta, esses resultados não foram acompanhados por um retorno significativo, levando a um índice de Sharpe inferior. Esse padrão vai ao encontro do arcabouço teórico basilar de Markowitz (1952), que propõe um trade-off entre risco e retorno, onde o investidor deve ponderar a relação risco-retorno ao selecionar sua carteira. Os resultados sugerem que a minimização do risco isoladamente não garante um desempenho superior, refletindo a necessidade de diversificação e balanceamento, conforme estabelecido por Markowitz.

Adicionalmente, o IFIX destacou-se como a estratégia com a maior volatilidade, porém, também apresentou o maior retorno e o índice de Sharpe mais elevado. Esse comportamento é consistente com a teoria do prêmio de risco, conforme discutido por Sharpe (1966), onde maior risco é geralmente associado a maior retorno esperado. O superior desempenho do IFIX pode ser parcialmente atribuído à sua maior diversificação, englobando 111 fundos, o que reduz o risco idiossincrático e possibilita capturar o crescimento dos setores mais dinâmicos, como fundos de shoppings e lajes corporativas, que responderam positivamente ao contexto econômico de 2022.

Uma análise análoga se verifica na estratégia de mínima variância, que detém a menor volatilidade e beta entre as estratégias consideradas. Contudo, evidencia-se também que essa estratégia apresenta o menor retorno e índice de Sharpe em comparação com as demais. Dessa forma, destaca-se uma superioridade das outras estratégias, corroborando a orientação de Markowitz (1952) de que os investidores devem buscar um equilíbrio entre esses dois aspectos.

Tabela 1 - Métricas de análise das carteiras e índice referência

	IFIX	SAC	DDPG	PPO	A2C	TD3	B&H	MINVAR
Retorno	**5,46**	-3,20	-3,35	-3,14	-3,12	-4,74	-3,15	-6,49
Retorno acumulado	**7,58**	-4,37	-4,57	-4,29	-4,26	-6,45	-4,30	-8,80
Volatilidade	12,04	6,84	6,11	6,43	6,21	6,66	6,44	**5,46**
Sharpe	**0,50**	-0,44	-0,52	-0,46	-0,48	-0,69	-0,46	-1,20
Beta	1	0,23	0,20	0,22	0,21	0,21	0,22	**0,16**

Fonte: Dados da pesquisa.

4.1 Transações realizadas

A análise das transações realizadas revela que o comportamento de compra e venda das estratégias de RL refletiu mudanças no ambiente macroeconômico e político. Observa-se, por exemplo, que o aumento expressivo na participação do HGBS11 a partir de junho de 2022 coincidiu com um crescimento substancial dos recebíveis do fundo. Esse comportamento sugere uma resposta adaptativa das estratégias a mudanças nos fundamentos do ativo, corroborando a hipótese de que as técnicas de RL são capazes de capturar variações de curto prazo, conforme discutido por Sutton e Barto (2018).

A composição da carteira pelos 26 FIIs mencionados e considerando 1%, assim como Santos et al. (2023) de custos de transação revelou movimentações notáveis de compra e venda em ativos específicos (HGBS11, KNIP11, HGLG11 e VRTA11), conforme evidenciado na Figura 1. Destaque para o fundo de shoppings HGBS11, que teve participação expressiva na carteira, enquanto os fundos de títulos e valores mobiliários, setores industriais e logística apresentaram uma relativa queda entre junho de 2022 e abril de 2023.

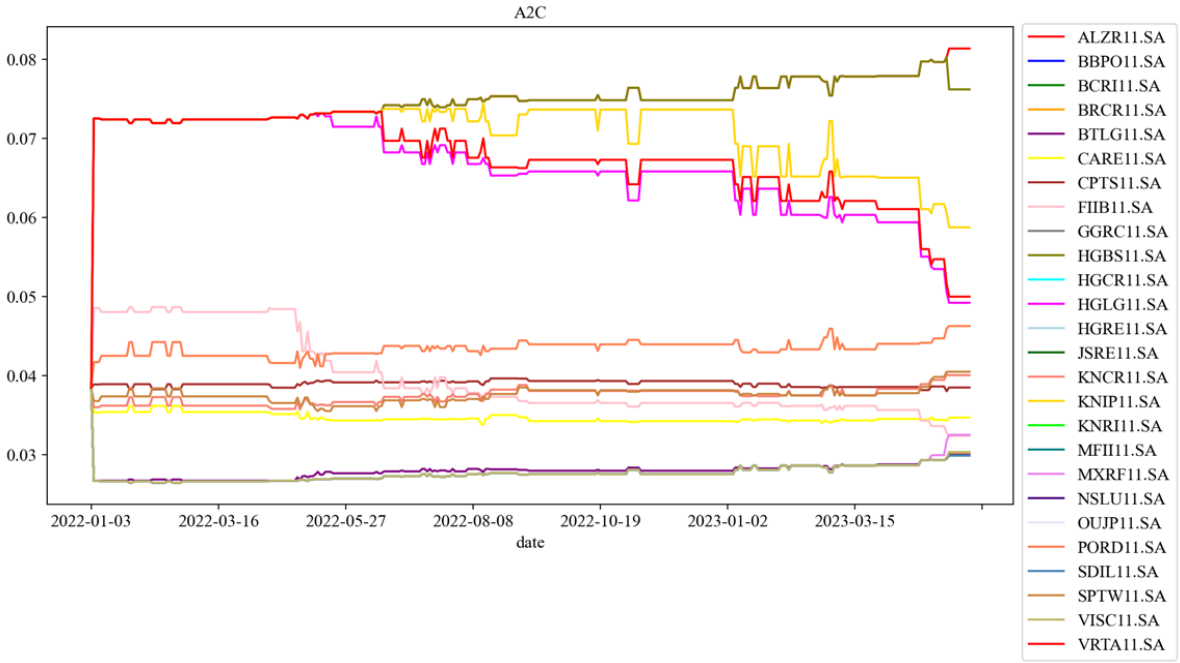
Do ponto de vista político-econômico, eventos como as reuniões do COPOM e do FED influenciaram diretamente a composição das carteiras. A postura mais cautelosa adotada pelo COPOM e a sinalização do FED sobre uma desaceleração no aumento das taxas de juros geraram um

impacto positivo na demanda por fundos de shoppings e títulos e valores mobiliários, refletindo a sensibilidade das estratégias a mudanças na política monetária. Esses resultados são consistentes com estudos de Chen et al. (1986), que destacam a relevância de fatores macroeconômicos na precificação de ativos financeiros.

É importante ressaltar o aumento significativo da participação do HGBS11 a partir de junho de 2022, possivelmente relacionado ao crescimento nos valores totais a receber desse fundo (aluguel, venda e outros), que passaram de cerca de R$ 15,3 milhões em abril de 2022 para R$ 34,1 milhões no mês seguinte, representando uma margem de recebíveis inédita nos últimos 5 anos.

Nota-se que tal resultado pode ser devido as técnicas propostas serem capazes de ajustar suas estratégias de alocação de ativos de acordo com as condições do mercado. Santos et al. (2023), notaram que a implementação dos custos de transação influencia na escolha dos investimentos, destacando a importância de considerar as condições de mercado ao desenvolver e avaliar estratégias de investimento.

Figura 1 - Gráfico do desempenho do A2C



Fonte: Dados da pesquisa.

Adicionalmente, do ponto de vista político-econômico, a mudança na composição da carteira pode ter sido influenciada por eventos como as reuniões do COPOM e do FED, próximas a abril de 2022. A postura mais cautelosa adotada pelo COPOM e a indicação de uma pausa no aumento da taxa de juros pelo FED podem ter impactado positivamente determinados setores do mercado financeiro, resultando em uma ampliação da participação do fundo de shoppings HGBS11 na carteira.

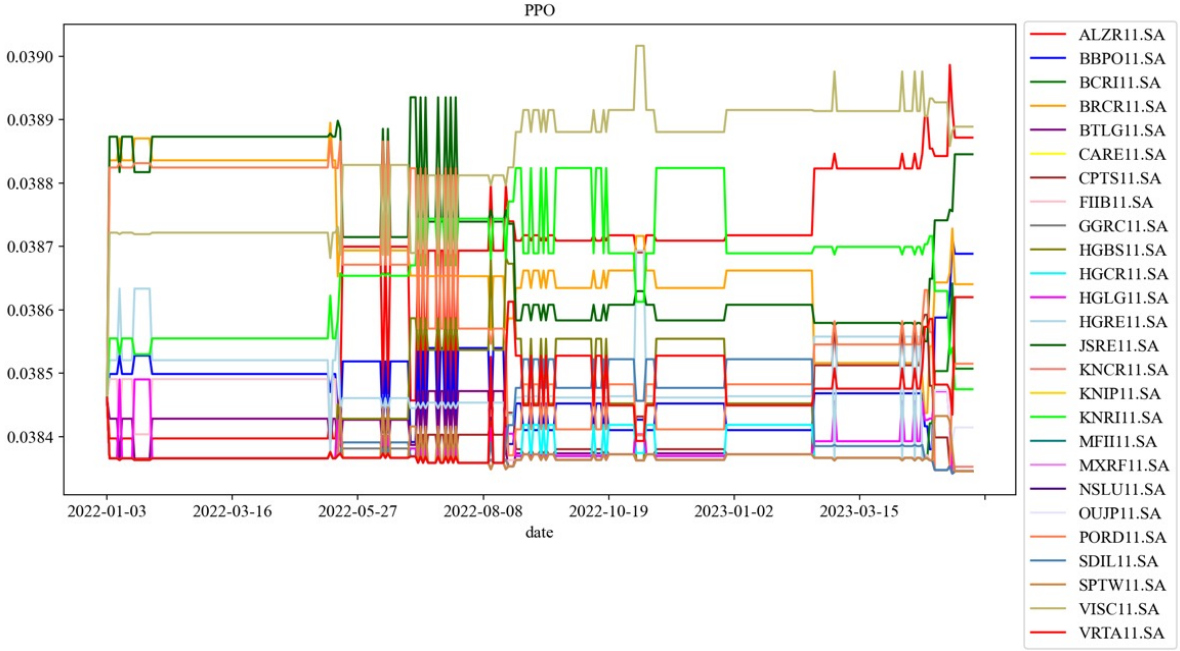
Paralelamente, a redução nos preços das commodities pode ter afetado negativamente os setores industriais e logísticos, resultando em uma diminuição na participação desses fundos, já que os valores a receber do HGLG11 caíram de uma média de R$ 200 milhões para R$ 124 milhões entre maio e setembro de 2022.

Além disso, essa queda foi captada pelas estratégias de RL, que diminuíram a exposição a ativos como HGLG11. Esse comportamento está alinhado com a literatura sobre modelagem financeira adaptativa (Lo, 2004), que sugere que mudanças nas condições de mercado levam a ajustes nas estratégias, visando minimizar perdas e capturar novas oportunidades de valorização.

O gráfico do algoritmo PPO (Figura 2), destaca uma maior atividade de compra e venda em

comparação com o A2C ao longo da pesquisa. Dois momentos notáveis incluem oscilações na carteira, coincidindo com os períodos entre junho e julho de 2022 e setembro e outubro de 2022. De acordo com Santos et al. (2023), a análise das alocações de ativos e a variação nas distribuições entre os modelos sugerem que os modelos de aprendizado por reforço, podem ter diferentes frequências de operação com base em suas estratégias. Modelos que adotam uma abordagem mais dinâmica e reativa às condições de mercado tendem a realizar mais operações.

Figura 2 - Gráfico do desempenho PPO



Fonte: Dados da pesquisa.

Em ambos os períodos analisados, houve alterações significativas nos fundos compostos por diferentes categorias, como fundos híbridos, lajes corporativas, títulos e valores mobiliários em papel, shoppings e industriais e logísticos. Apesar da consistência nas categorias, os ativos específicos variaram entre os dois momentos. Essas variações podem estar associadas às decisões das reuniões do COPOM e do FED, uma vez que os tipos de fundos afetados foram os mesmos para os dois algoritmos estudados.

Tanto no Momento I quanto no Momento II, a postura cautelosa do COPOM ao manter a taxa SELIC em 13,75% e a indicação do FED de uma desaceleração no aumento das taxas de juros, com um acréscimo de 0,25%, geraram otimismo no mercado financeiro. Isso resultou em uma maior demanda por fundos de shoppings e títulos e valores mobiliários. A alta inflação em março, medida pelo IPCA (1,62%), e o IPCA-15 de abril (1,73%), abaixo das expectativas de mercado, podem ter impactado os fundos de lajes corporativas e industriais e logísticos, mais sensíveis às flutuações de preços. O ciclo de aperto monetário nos Estados Unidos também pode ter influenciado, tornando a taxa de juros brasileira mais competitiva para atrair investimentos.

Tais análises estão de acordo com a literatura cientifica, pois segundo Scolese et al. (2015), os retornos dos FIIs são influenciados tanto por fatores de renda fixa, como as taxas de juros e a inflação, quanto por fatores de renda variável, como o Ibovespa. Além de Oliveira e Milani (2020) que verificaram um maior impacto do índice ibovespa no desempenho dos fundos entre 2013 e 2017 no Brasil.

Os gráficos dos algoritmos DDPG, SAC e TD3 não apresentaram padrões distintos em relação aos dois gráficos analisados anteriormente. Além disso, registraram um menor número de operações

comparado aos métodos A2C e PPO.

4.2 Retorno Acumulado

Os resultados apresentados na Figura 3 ilustram a evolução do retorno acumulado das estratégias de RL e comparáveis. A estratégia SAC destaca-se positivamente entre abril e dezembro de 2022, atribuída à concentração de investimentos em VRTA11. Esse desempenho reforça a ideia de que estratégias algorítmicas podem identificar e explorar momentos específicos de valorização de ativos, conforme discutido por Sun et al. (2023). No entanto, a partir de julho de 2022, o índice IFIX supera as carteiras elaboradas, sugerindo que a maior diversificação proporcionada pelo índice foi um fator-chave para mitigar a volatilidade e capturar retornos superiores.

Esse achado vai ao encontro do modelo de seleção de carteiras de Markowitz (1952), que defende que uma maior diversificação pode reduzir o risco sem sacrificar o retorno. No caso das estratégias de RL, a limitação a um conjunto restrito de ativos pode ter levado a um desempenho inferior ao IFIX. Embora as estratégias de RL possuam potencial para identificar padrões complexos no curto prazo, sua performance é sensível à composição da carteira e ao período de análise, aspectos discutidos por Sun et al. (2023) em relação à modelagem de portfólios dinâmicos.

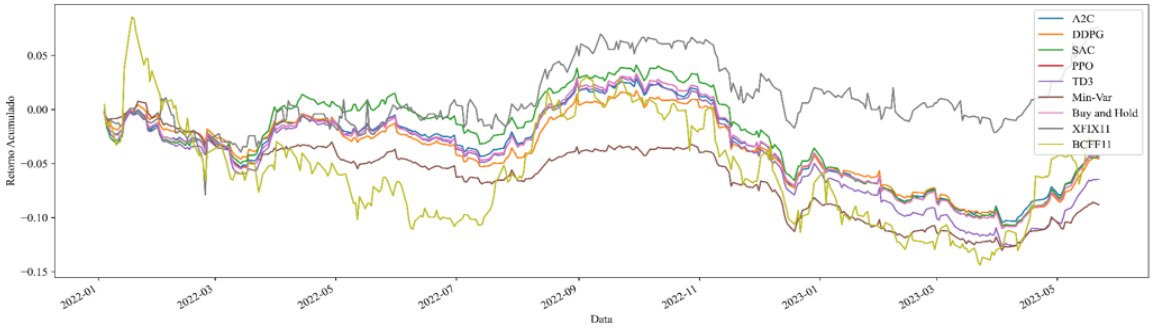
A carteira dos algoritmos de RL inicialmente acompanhou o desempenho do XIFIX11, mas a partir de julho de 2022, o índice superou as carteiras elaboradas e outras estratégias do estudo, obtendo resultados semelhantes ao de Iório et al. (2015). Essa diferença pode ser explicada pela maior diversificação do IFIX, composto por 111 fundos.

Por fim, a comparação entre as estratégias B&H e de mínima variância com as de RL evidencia que as carteiras algorítmicas priorizam a compreensão de padrões e maximização de retornos, enquanto a mínima variância busca minimizar o risco, conforme a abordagem clássica de Harry Markowitz. Essa dicotomia de objetivos resultou em perfis de risco e retornos divergentes, favorecendo as estratégias de RL em termos de retorno absoluto, mas mantendo um desempenho inferior ao IFIX, conforme preconizado na literatura sobre mercados adaptativos e eficiências relativas (Lo, 2004).

Além disso, as estratégias de mínima variância e B&H não superaram os algoritmos, apresentando desempenho inferior, resultado oposto ao de Iório et al. (2015), que não conseguiu obter resultados superiores a técnicas tradicionais. Isso decorre da natureza divergente dos objetivos e estratégias, onde as carteiras algorítmicas, como DDPG, SAC, TD3, A2C e PPO, priorizam a maximização do retorno. Assim, as distintas abordagens resultaram em perfis de risco e desempenhos divergentes, favorecendo as carteiras algorítmicas.

Essas conclusões reforçam a necessidade de um balanceamento cuidadoso entre os objetivos de risco e retorno, bem como uma revisão contínua das estratégias em função das mudanças nos regimes de mercado, conforme destacado por Santos et al. (2023) em sua análise sobre gestão de portfólios de investimento em ambientes dinâmicos.

Figura 3 - Gráfico do Retorno Acumulado



Fonte: Dados da pesquisa

5 CONCLUSÃO

O estudo analisou o desempenho de algoritmos de aprendizado por reforço na composição de carteiras de Fundos de Investimento Imobiliário (FIIs). Os resultados indicam que os algoritmos A2C, DDPG, PPO, SAC, TD3 e a estratégia B&H apresentaram desempenhos similares e superiores às estratégias de mínima variância. Além disso, eventos econômicos e políticos, como reuniões do COPOM e do FED, junto com variações nos preços das commodities e na inflação, podem ter influenciado o comportamento dos fundos na carteira.

O estudo segue a Hipótese de Mercado Eficiente, divergindo da Hipótese de Mercado Adaptativo, uma vez que não foi possível superar os resultados do IFIX. Em relação as limitações, incluem a focalização em um período e conjunto específicos de FIIs, além da não consideração de outros fatores que poderiam influenciar o desempenho dos fundos na carteira.

As implicações estratégicas dos procedimentos adotados e dos achados são significativas para investidores e gestores de carteiras. Primeiro, o uso de RL na otimização de carteiras demonstrou sua eficácia ao superar estratégias tradicionais, proporcionando um novo caminho para quem busca maximizar retornos com menor risco. No entanto, o fato de que as carteiras elaboradas pelos algoritmos não superaram o IFIX sugere que, em mercados como o de Fundos de Investimento Imobiliário (FIIs), pode haver uma eficiência que limita ganhos extraordinários, alinhando-se à HME de Fama (1970).

Este estudo destaca duas limitações significativas: a restrição na quantidade de Fundos de Investimento Imobiliário elegíveis para a pesquisa e a limitação temporal dos dados coletados. Para futuras investigações, sugere-se a expansão da amostra de FIIs e o aumento do período de análise. Além disso, explorar estratégias de otimização de hiperparâmetros, algo ainda desafiador nesse contexto, para ajustar a configuração dos algoritmos de RL e potencialmente aprimorar o desempenho das carteiras. Essa abordagem permitiria uma investigação mais aprofundada sobre o impacto de diferentes combinações de hiperparâmetros na estabilidade e nos retornos das estratégias desenvolvidas.

REFERÊNCIAS

ASSAF NETO, Alexandre. **Mercado Financeiro**. 13. ed. São Paulo: Atlas, 2019.

ASSAF NETO, Alexandre; LIMA, Fabiano Guasti; DE ARAÚJO, Adriana Maria Procópio. Uma proposta metodológica para o cálculo do custo de capital no Brasil. **Revista de Administração**-RAUSP, v. 43, n. 1, p. 72-83, 2008.

BURNHAM, Terence C. Toward a neo-Darwinian synthesis of neoclassical and behavioral economics. **Journal of Economic Behavior & Organization**, v. 90, p. S113-S127, 2013.

B3. **Manual de Definições e Procedimentos dos índices da B3** São Paulo, fevereiro de 2018. Disponível em: https://www.b3.com.br/data/files/CA/A5/9F/28/14F35810F534EB48AC094EA8/Manual%20de%20defini%C3%A7%C3%B5es%20e%20procedimentos%20de%20%C3%8Dndices-PT.pdf. Acesso em: 9 jul. 2023.

CASTELLO BRANCO, Carlos Eduardo; MONTEIRO, Eliane de Mello Alves Rebouças. Estudo sobre a indústria de fundos de investimentos imobiliários no Brasil. **Revista do BNDES**, v. 10, n. 20, p. 261-295, 2003.

CHAN, Su Han; ERICKSON, John; WANG, Ko. **Real Estate Investment Trusts: Structure, performance, and investment opportunities**. Financial Management Association Survey and Synthesis, 2003.

CHEN, K. C.; DORSEY, R. E.; KWON, C. The effect of market structure on asset pricing. **Journal of Finance**, v. 41, n. 1, p. 207-224, 1986.

CVM. **Texto Integral da Instrução CVM Nº 205, de 14 de Janeiro de 1994, com Alterações Introduzidas pelas Instruções. CVM NoS 389/03, 418/05 E 455/07.** Janeiro de 1994. Disponível em: https://conteudo.cvm.gov.br/export/sites/cvm/legislacao/instrucoes/anexos/200/inst205consolid.pdf. Acesso em: 23 set. 2024.

DULAC-ARNOLD, Gabriel et al. An empirical investigation of the challenges of real-world reinforcement learning. **arXiv preprint arXiv:2003.11881**, 2020.

FAMA, Eugene F. Efficient capital markets: A review of theory and empirical work. **The Journal of Finance**, v. 25, n. 2, p. 383-417, 1970.

FANG, Yunzhe; LIU, Xiao-Yang; YANG, Hongyang. Practical machine learning approach to capture the scholar data driven alpha in AI industry. In: **2019 IEEE International Conference on Big Data (Big Data).** IEEE, 2019. p. 2230-2239.

FENG, Zhilan; PRICE, S. McKay; SIRMANS, C. An overview of equity real estate investment trusts (REITs): 1993–2009. **Journal of Real Estate Literature**, v. 19, n. 2, p. 307-343, 2011.

FUJIMOTO, Scott; HOOF, Herke; MEGER, David. Addressing function approximation error in actor-critic methods. In: **International conference on machine learning**. PMLR, 2018. p. 1587-1596.

GELTNER, David et al. **Commercial real estate analysis and investments**. Cincinnati, OH: South-western, 2001.

HAARNOJA, Tuomas et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: **International conference on machine learning**. PMLR, 2018. p. 1861-1870.

HAMBLY, Ben; XU, Renyuan; YANG, Huining. Recent advances in reinforcement learning in finance. **Mathematical Finance**, v. 33, n. 3, p. 437-503, 2023.

HAN, Seungyul; SUNG, Youngchul. A max-min entropy framework for reinforcement learning. **Advances in Neural Information Processing Systems**, v. 34, p. 25732-25745, 2021.

IÓRIO, F.R.; LUCCHESI, E.P.; IIZUKA, E.S. 2015. **Análise do desempenho de carteiras de fundos de investimento imobiliário negociados na BM&FBOVESPA entre 2011 e 2013**. In: Seminários em Administração, XVIII, São Paulo, p. 1-14. 2015.

LILLICRAP, Timothy P. et al. Continuous control with deep reinforcement learning. **arXiv preprint arXiv:1509.02971,** 2015.

LIOW, Kim Hiang; HUANG, Yuting; SONG, Jeonseop. Relationship between the United States housing and stock markets: some evidence from wavelet analysis. **The North American Journal of Economics and Finance**, v. 50, p. 101033, 2019.

LIOW, Kim Hiang; HUANG, Yuting; SONG, Jeongseop. Who influences the Asian–Pacific real estate markets: the US, Japan or China?. **China & World Economy**, v. 27, n. 6, p. 50-78, 2019.

LIU, Xiao-Yang et al. FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance. **arXiv preprint arXiv:2011.09607**, 2020.

LO, Andrew W. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective**.** **Journal of Portfolio Management**, v. 30, n. 1, p. 15-29, 2004.

MA, Xiaoteng et al. Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning. **arXiv preprint arXiv:2004.14547**, 2020.

MARKOWITZ, Harry. Portfolio selection. **Journal of Finance**, v.7, n.1, p. 77-91, 1952.

MNIH, Volodymyr et al. Asynchronous methods for deep reinforcement learning. In: **International conference on machine learning**. PMLR, p. 1928-1937, 2016.

OLIVEIRA, Janaína Morais de; MILANI, Bruno. Variáveis que explicam o retorno dos fundos imobiliários brasileiros. **Revista Visão: Gestão Organizacional**, v. 9, n. 1, p. 17-33, 2020.

OSHINGBESAN, Adebayo et al. Model-Free Reinforcement Learning for Asset Allocation. **arXiv preprint arXiv:2209.10458**, 2022.

SANTOS, Gustavo Carvalho et al. Management of investment portfolios employing reinforcement learning. **PeerJ Computer Science**, v. 9, p. e1695, 2023.

SCOLESE, Daniel et al. Análise de estilo de fundos imobiliários no Brasil. **Revista de Contabilidade e Organizações**, v. 9, n. 23, p. 24-35, 2015.

SCHULMAN, John et al. Proximal policy optimization algorithms. **arXiv preprint arXiv:1707.06347,** 2017.

SCHULMAN, John et al. Trust region policy optimization. In: **International conference on machine learning**. PMLR, 2015. p. 1889-1897.

SHARPE, William F. A simplified model for portfolio analysis. **Management science**, v. 9, n. 2, p. 277-293, 1963.

SHARPE, William F. Capital asset prices: A theory of market equilibrium under conditions of risk. **The Journal of finance**, v. 19, n. 3, p. 425-442, 1964.

SHARPE, William F. Mutual fund performance. **The Journal of Business**, v. 39, n. 1, p. 119-138, 1966.

SUN, Shuo; WANG, Rundong; AN, Bo. Reinforcement learning for quantitative trading. **ACM Transactions on Intelligent Systems and Technology**, v. 14, n. 3, p. 1-29, 2023.

SUTTON, Richard S. et al. Policy gradient methods for reinforcement learning with function approximation. **Advances in Neural Information Processing Systems**, p. 1057-1063, 2000.

SUTTON, R. S.; BARTO, A. G. (2018). Reinforcement Learning: An Introduction. **A Bradford Book,** 2ª ed. Boston: MIT Press.

WANG, Ziyu et al. Sample efficient actor-critic with experience replay. **arXiv preprint arXiv:1611.01224**, 2016.

YANG, Hongyang et al. Deep reinforcement learning for automated stock trading: An ensemble strategy. In: **Proceedings of the first ACM International Conference on AI in Finance**. p. 1-8. 2020.

ZHANG, Zihao; ZOHREN, Stefan; ROBERTS, Stephen. Deep reinforcement learning for trading. **The Journal of Financial Data Science**, v. 2, n. 2, p. 25-40, 2020.