SemeAd 2023
XXVI Seminários em Administração

07, 08, 09 e 10 de novembro de 2023
ISSN 2177-3866

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

JULIA PINHEIRO BARBOZA
UNIVERSIDADE FEDERAL DE UBERLÂNDIA (UFU)

GUSTAVO CARVALHO SANTOS
UNIVERSIDADE FEDERAL DE UBERLÂNDIA (UFU)

DANIEL VITOR TARTARI GARRUTI
UNIVERSIDADE FEDERAL DE UBERLÂNDIA (UFU)

FLAVIO LUIZ DE MORAES BARBOZA
UNIVERSIDADE FEDERAL DE UBERLÂNDIA (UFU)

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

1. Introdução

Markowitz (1952) desenvolveu a teoria moderna de portfólio, propondo a diversificação de investimentos para reduzir riscos. Fama (1970) propôs a Hipótese de Mercado Eficiente, que sugere que os preços dos ativos refletem todas as informações disponíveis, enquanto Lo (2004) propôs a Hipótese de Mercado Adaptativo que indica que os mercados são eficientes, mas podem se tornar ineficientes em momentos de incerteza ou mudanças rápidas. Durante o estudo, ambas as teorias serão exploradas e correlacionadas com a gestão de portfólio proposta, permitindo escolher qual delas fundamenta de maneira mais coerente a estratégia adotada.

A proposição de Fundo de Investimento Imobiliário (FII) foi estabelecida nos Estados Unidos na década de 1880, mas começou a prosperar na década de 1960, sob o formato de *Real estate investment trust* (Reit), quando o Congresso aprovou uma legislação que concedia benefícios fiscais similares aos dos *mutual funds* (Branco e Monteiro, 2003). Desde então, os regulamentos dos REITs passaram por diversas mudanças, incluindo a Lei de Reforma Tributária de 1986 (TRA), a carta particular do IRS sobre a oferta pública inicial (IPO) da Taubman Centers Inc. de 1992, a Lei de Orçamento e Reconciliação Omnibus de 1993 (OBRA) e a Lei de Modernização do REIT de 1999 (RMA), o que teve um impacto significativo no tamanho, natureza e composição da indústria (Feng, Price e Sirmans, 2011; GELTNER et al., 2001; Chan, Erickson e Wang, 2003). Os REITs cresceram em valor de mercado, passando de apenas US$ 26 bilhões em 1993 para mais de US$ 400 bilhões em 2006 (Feng, Price e Sirmans, 2011).

Apesar disto, os Fundos Imobiliários adquiriram ímpeto no Brasil na década de 1990, quando foram estabelecidos legalmente em Junho de 1993, pela Lei 8.668, sancionada pelo Presidente da República. Esta lei emergiu a constituição e o regime tributário dos Fundos de Investimento Imobiliário e dos Fundos de Investimento nas Cadeias Produtivas Agroindustriais (Fiagro), além de definir as alíquotas e obrigações fiscais das instituições administradoras, contribuindo para o desenvolvimento dos setores imobiliário e agroindustrial do país. A Instrução 205 da Comissão de Valores Mobiliários (CVM), de 1994, também foi de suma importância, regulamentando a constituição, o funcionamento e a administração desses fundos, garantindo transparência e segurança aos investidores, definindo as obrigações e responsabilidades das instituições administradoras, bem como os critérios para a avaliação dos ativos do fundo e negociação das quotas.

Sob esse contexto e em conformidade com informações publicadas pela [B]³ (boletim público Nº 71, 2018), o mercado de fundos imobiliários no Brasil em 2018 representava 368 fundos totais (registrados pela CVM), sendo 154 listados pela bolsa, os quais possuíam um patrimônio líquido total de R$ 80,80 bilhões – registrados totais - e o valor de mercado de R$ 45 bilhões, no que tange aos fundos listados pela Bovespa, naquele ano. Já em 2023, também em divulgações públicas da [B]³ (abril, 2023), o número de fundos totais apresentou 819, sendo 483 listados pela bolsa, com o patrimônio líquido total de mais de R$ 200 bilhões e o valor de mercado de R$ 138 milhões. Sendo assim, visto o tamanho deste mercado, o ritmo de crescimento nos últimos anos e seu potencial de negociações, os fundos de investimento imobiliários foram selecionados como base de pesquisa para o presente estudo.

Diante dessa perspectiva, o presente estudo procurou apresentar resultados de desempenho de uma carteira elaborada durante a pesquisa, de modo a apresentar o desempenho de carteiras elaboradas com algoritmos de aprendizado por reforço, comparando-o com estratégias tradicionais e avaliando o impacto de eventos econômicos e políticos. Como resultados, foi

observado que o método SAC se sobressaiu positivamente, enquanto o índice IFIX superou as carteiras elaboradas pelos algoritmos após Julho de 2022. As estratégias de mínima variância e *Buy and Hold* não conseguiram superar os algoritmos, obtendo um desempenho inferior e semelhante, respectivamente.

Este artigo está dividido em cinco seções. Após esta introdução, apresenta-se o referencial teórico, incluindo a Teoria de Carteiras e o uso do Reinforcement Learning para gestão de carteiras. Em seguida, são descritos os procedimentos metodológicos, incluindo a seleção dos fundos imobiliários pertencentes as carteiras elaboradas e a descrição dos algoritmos utilizados da FinRL. Na seção 4, são apresentados os resultados obtidos com os diferentes algoritmos. Por fim, na seção 5, a pauta é a conclusão do trabalho.

2. Referencial teórico

2.1 Teoria de Carteiras

A teoria de carteiras, proposta por Harry Markowitz em 1952, é uma das mais importantes teorias da área de finanças. Segundo Markowitz (1952), a teoria de carteiras pressupõe que os investidores são racionais e buscam maximizar a utilidade esperada do seu patrimônio. Isso significa que os investidores levam em conta tanto o retorno quanto o risco dos ativos ao tomar suas decisões de investimento. Além disso, a teoria pressupõe que os investidores são avessos ao risco e preferem carteiras com menor risco para um dado nível de retorno. A relação risco-retorno é fundamental na teoria de carteiras, pois permite ao investidor avaliar o desempenho da carteira em termos de retorno e risco. Segundo Markowitz (1952), a relação risco-retorno é positiva, ou seja, quanto maior o risco, maior o retorno esperado. No entanto, o investidor pode escolher uma carteira que minimize o risco para um dado nível de retorno ou maximize o retorno para um dado nível de risco.

Dessa forma, a diversificação é uma estratégia importante para reduzir o risco total da carteira, pois permite ao investidor combinar ativos com baixa correlação entre si, sendo uma estratégia importante na teoria de carteiras, pois reduz o risco total da carteira sem reduzir o retorno esperado. Segundo Markowitz (1952), a diversificação é possível porque os ativos têm correlação negativa entre si, ou seja, quando um ativo está em alta, outro está em baixa. No que diz respeito à avaliação de risco e desempenho das carteiras, Markowitz (1952) propôs o conceito de risco como a variância dos retornos, que é uma medida da dispersão dos retornos em torno da média. Essa medida captura o grau de incerteza sobre os retornos futuros e permite comparar diferentes ativos ou carteiras em termos de risco.

Além disso, Markowitz (1952) propôs o conceito de relação risco-retorno, que indica quanto retorno adicional o investidor pode obter ao assumir um risco adicional. Essa relação é fundamental para avaliar o desempenho das carteiras, pois permite verificar se o retorno obtido compensa o risco assumido. Sendo assim, em relação aos fundos de investimento imobiliários (FIIs) ou *Real estate investment trust* (REITs), é possível aplicar a teoria de carteiras para selecionar uma carteira ótima desses ativos. O modelo de Markowitz pode ser aplicado a qualquer tipo de ativo, incluindo os imobiliários, desde que se conheça o retorno esperado e a variância dos retornos de cada ativo, bem como a covariância entre os ativos. Assim, o modelo de Markowitz pode auxiliar na gestão de carteiras de fundos imobiliários, permitindo ao investidor diversificar seus investimentos e reduzir o risco total da carteira. Outras medidas de risco, como o Coeficiente Beta de Sharpe (1963) e o Índice de Sharpe (1964), são usadas para avaliar o desempenho ajustado ao risco de um portfólio.

Segundo Assaf Neto (2019), os Fundos de Investimento apresentam uma relação direta entre risco e retorno. Quanto maior a possibilidade de rendimento de um Fundo, maior também será o risco incorrido pelo investidor. Em contrapartida, fundos que oferecem maior segurança a

seus cotistas costumam apresentar um retorno menor. A escolha da relação risco-retorno que mais se adequa ao investidor é uma decisão a ser realizada por ele, definida por sua aversão ao risco. Nesse sentido, os principais tipos de risco presentes nos Fundos de Investimentos são o risco de crédito, o risco de mercado, o risco de liquidez e o risco sistêmico.

2.2 Hipóteses do Mercado Financeiro

A Hipótese dos Mercados Eficientes, proposta por Fama (1970), parte do pressuposto de que os preços dos ativos financeiros refletem todas as informações disponíveis. A hipótese pode ser classificada em três formas: fraca, semi-forte e forte. A forma fraca sustenta que o mercado reflete todas as informações públicas disponíveis. A forma semi-forte engloba a forma fraca e sugere que as novas informações são absorvidas instantaneamente pelo mercado. A forma forte abrange os dois outros formatos e afirma que os preços refletem todo o tipo de informação, tanto pública quanto privada.

Lo (2004) estabelece a Hipótese de Mercados Adaptativos (HMA), fundamentada sob a perspectiva da análise de mercados e pode ser entendida como uma evolução da Hipótese dos Mercados Eficientes de Fama. A HMA refere-se a princípios evolucionários, voltados para a economia, onde leis biológicas como seleção natural, adaptação, mutação e aprendizado orientam quais estratégias e heurísticas de tomada de decisão são as mais apropriadas. A HMA concilia a estrutura neoclássica da HME com o comportamento não ótimo do agente, considerando novas estruturas de tomada de decisão financeira pelo investidor, como aprendizado, adaptação e vieses comportamentais (Burnham, 2013). Sendo assim, de acordo com a HMA, é possível prever os preços dos ativos do mercado financeiro.

2.3 Trabalhos relacionados

Scolese et al. (2015) foram responsáveis por estruturar uma pesquisa, que investigou o retorno dos fundos de investimentos imobiliários (FIIs) no Brasil, buscando identificar seu estilo e, consequentemente, seu comportamento frente aos índices do mercado financeiro brasileiro do segmento de renda fixa, de renda variável e do segmento imobiliário para o período de 2011 a 2015. Os resultados obtidos indicaram que os retornos dos FIIs acompanham de forma mais pronunciada os juros prefixados e os retornos do mercado imobiliário.

Um estudo amplo sobre os REITs de capital aberto foi realizado por Feng, Price e Sirmans (2011), no qual apresentaram o crescimento e consolidação da indústria, as mudanças no foco do tipo de propriedade, o aumento da propriedade institucional e o aumento do uso de parcerias operacionais. Além disso, mostram variações nas métricas de desempenho contábil, aumentos na alavancagem e flutuações no fluxo de caixa. O artigo apresenta um panorama do estado da arte dos REITs de capital aberto durante a época estudada.

Iorio e Lucchesi (2014) conduziram uma análise comparativa do desempenho de três portfólios de Fundos de Investimento Imobiliário (FIIs): (a) um portfólio composto por 10 FIIs construído com base na teoria proposta por Markowitz (1952), (b) um portfólio composto por 10 FIIs com pesos iguais (portfólio simplificado) e (c) o índice IFIX, durante o período de 2011 a 2013. Os pesquisadores não encontraram diferenças expressivas entre os retornos dos três portfólios. Todavia, a avaliação do desempenho considerando a combinação de risco e retorno aponta que o IFIX apresentou um desempenho superior em relação ao portfólio teórico, que, por sua vez, obteve um desempenho superior ao portfólio simplificado.

Um estudo de Yang et al. (2020) propôs uma estratégia de conjunto para negociação automatizada de ações utilizando o aprendizado por reforço profundo. A estratégia integra três algoritmos baseados em ator-crítico: *Proximal Policy Optimization* (PPO), *Advantage Actor Critic* (A2C) e *Deep Deterministic Policy Gradient* (DDPG). A pesquisa teve por objetivo testá-

los em conjunto para superar os algoritmos em seu formato individual e outras linhas de base em termos de retorno ajustado ao risco (Índice de Sharpe). Desse modo, a estratégia conjunta superou os algoritmos. Nesta pesquisa, esta estratégia de combinação será adotada, com os mesmos algoritmos do estudo de Yang et al. (2020) e com a adição de outros dois algoritmos – os quais serão explorados posteriormente.

Nesse contexto, o artigo de Sun, Wang e An (2023) contribuiu com uma pesquisa abrangente acerca dos esforços de pesquisa, em métodos baseados em aprendizado por reforço, para tarefas de negociação quantitativa (*Quantitative Trading*). Dessa forma, forneceu uma visão geral do estado da arte no que que se refere ao uso de modelos matemáticos – internos a biblioteca FinRL - e técnicas baseadas em dados para analisar o mercado financeiro e identificar oportunidades de investimento, assim como destacou desafios e direções futuras de pesquisa nesse campo.

2.4 *Reinforcement Learning* para gestão de carteiras

O método de *Reinforcement Learning* ou Aprendizado por Reforço (RL) é um subcampo da aprendizagem de máquina que envolve treinar um agente para aprender uma série de ações que maximizam uma recompensa cumulativa, o que pode incluir negociação, como visto no artigo "Aprendizado Profundo por Reforço para Negociação Automatizada de Ações: Uma Estratégia de Conjunto" (YANG et al., 2020).

O artigo apresentado por Yang et al. na ICAIF 2020 descreve uma estratégia de conjunto para negociação automatizada de ações usando aprendizado profundo por reforço. A estratégia combina várias técnicas de RL para criar um modelo robusto capaz de lidar com as incertezas do mercado financeiro. Os autores discutem como o uso de múltiplos agentes de RL pode aumentar a capacidade do modelo de lidar com diferentes condições de mercado e apresentam resultados experimentais que demonstram a eficácia da abordagem.

Ademais, a pesquisa abrangente sobre esforços de pesquisa em métodos baseados em RL para Negociação Quantitativa, realizada por Shuo Sun, Rundong Wang e Bo An (SUN; WANG; AN, 2023) fornece uma visão geral dos esforços de pesquisa em métodos baseados em RL para Negociação Quantitativa. Os autores discutem os desafios e oportunidades na aplicação do RL ao campo da negociação e apresentam uma análise detalhada das principais abordagens e técnicas utilizadas. Eles também discutem como o RL pode ser usado para desenvolver estratégias de negociação adaptativas que podem se ajustar às mudanças nas condições do mercado.

3. Procedimentos Metodológicos

3.1 Fundos imobiliários pertencentes ao IFIX

Segundo a [B]³, a metodologia adotada pelo Índice de Fundos de Investimentos Imobiliários (IFIX) fundamenta-se na construção de uma carteira teórica de ativos, cuja seleção segue critérios estabelecidos. Os procedimentos e regras inerentes ao IFIX são detalhados no Manual de Definições e Procedimentos dos Índices da B3. Nessa fundamentação, o IFIX é classificado como um índice de retorno total e seu objetivo primordial consiste em desempenhar a função de indicador representativo do desempenho médio das cotações dos fundos imobiliários negociados nos mercados de bolsa e balcão organizado da B3. O processo de seleção dos fundos que integram o IFIX e sua orientação por critérios específicos, além de serem compostos por fundos listados na bolsa, motivou o início da segmentação da amostra de ativos elegíveis para o presente estudo. Neste primeiro momento a amostra representava 111 fundos.

3.1.2 Fundos imobiliários com, no mínimo, 5 anos de dados e exclusão Fundos de Fundos

A seleção de fundos imobiliários com um histórico de, no mínimo, 5 anos de dados está intrinsecamente relacionada à aplicação do método de Aprendizado por Reforço no contexto da

inteligência artificial. Tal exigência é motivada pela necessidade de uma análise histórica abrangente e sólida, permitindo ao algoritmo aprender e identificar padrões de desempenho dos fundos ao longo do tempo. Ao utilizar um período mais extenso, o Aprendizado por Reforço pode capturar uma quantidade significativa de informações passadas, aumentando a precisão e a confiabilidade da análise. Essa abordagem também minimiza a influência de flutuações de curto prazo, proporcionando uma visão mais estável e consistente do desempenho dos fundos imobiliários.

Além disso, a escolha pela exclusão de FOFs (Fundos de Fundos) ocorreu pela disposição e facilidade do alcance de dados em fundos que não incluíssem outros fundos em seu portfólio, já que o estudo teve por objetivo analisar uma carteira estabelecida a partir da amostra final de fundos, com características específicas, das quais os FOFs não atenderiam por serem compostos por um conjunto de ativos com atributos distintos e de difícil análise. Com a adição destes critérios a amostra foi reduzida para 40 fundos elegíveis para a pesquisa.

3.1.3 Fundos imobiliários com alto volume de negociações

A escolha de ativos líquidos para compor o índice IFIX está relacionada à preferência por ativos facilmente negociáveis. A liquidez é crucial na análise e acompanhamento desses fundos, permitindo uma formação de preços mais eficiente. Essa característica é essencial para o sucesso do algoritmo de Aprendizado por Reforço, que utiliza o alto volume de negociações para aprender padrões e identificar tendências (OSHINGBESAN et al., 2022; ZHANG; ZOHREN; ROBERTS, 2019). A abundância de dados fornecidos pelos ativos líquidos auxilia o algoritmo a tomar decisões mais embasadas (HAMBLY; XU; YANG, 2023). Ativos líquidos permitem que o agente execute suas estratégias de compra e venda de maneira eficiente. Dos fundos elegíveis, 26 fundos possuem o perfil de alto volume de negociações e foram selecionados para a amostra final.

A amostra final é composta pelos fundos a seguir:

TABELA 1 - Tabela de 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: Elaboração própria.

3.2 Coeficiente Beta

O conceito de coeficiente beta foi desenvolvido por William F. Sharpe, John Lintner e Jan Mossin, em artigos publicados na década de 1960. Sharpe (1963) propôs um modelo simplificado para análise de carteiras, baseado na hipótese de que os investidores são avessos ao risco e buscam maximizar a utilidade esperada de seus retornos. Sharpe definiu o coeficiente beta como a razão entre a covariância entre o retorno da carteira e o retorno do mercado e a variância do retorno do mercado. Sharpe mostrou que, em equilíbrio, o retorno esperado de uma carteira é uma função linear do seu coeficiente beta.

Lintner (1965) e Mossin (1966) estenderam o modelo de Sharpe para o caso de ativos individuais, assumindo que os investidores podem aplicar e tomar emprestado à taxa livre de risco. Lintner e Mossin demonstraram que, em equilíbrio, o retorno esperado de um ativo é igual ao retorno da taxa livre de risco mais um prêmio pelo risco sistemático, medido pelo produto entre o coeficiente beta do ativo e o prêmio pelo risco do mercado. Essa relação é conhecida como CAPM (*Capital Asset Pricing Model*).

Conforme Sharpe (1964), a fórmula para o cálculo do coeficiente beta é a seguinte:

$\beta = Cov\ (Ri, Rm) / Var\ (Rm)$

Onde: β é o coeficiente beta; Cov (Ri, Rm) é a covariância entre o retorno do ativo i e o retorno do mercado; Var (Rm) é a variância do retorno do mercado.

Segundo (Assaf Neto; Lima; Araújo, 2008), a carteira de mercado é a mais diversificada e contém apenas risco sistemático, quando o beta é igual a 1,0. Ativos com beta igual a 1,0 têm retorno igual ao retorno médio da carteira de mercado. Ativos com beta maior que 1,0 têm risco maior e maior expectativa de retorno. Ativos com beta menor que 1,0 têm risco menor e menor expectativa de retorno.

3.3 Índice de Sharpe

O índice de Sharpe (1964) é uma ferramenta que possibilita aferir o desempenho de um investimento em face do seu risco. Ele mensura qual é a relação entre o retorno excedente ao ativo isento de risco e sua volatilidade. O índice de Sharpe é concebido para auxiliar os investidores a compreender o retorno potencial de um investimento, em confronto com seu risco. Quanto maior o índice Sharpe, mais atrativo é o retorno ajustado ao risco, sendo assim, pode ser empregado para calcular o desempenho pretérito ou o desempenho esperado no futuro, estimando os números a serem preenchidos em cada campo da fórmula.

O índice de Sharpe é estabelecido pela seguinte equação:

Índice Sharpe = (Rp – Rf)/OP

Onde: Rp = Retorno do portfólio ou do ativo; Rf = *Risk Free Ratio*, que é a taxa livre de risco, ou seja, no âmbito brasileiro correponde a taxa Selic; OP = É o desvio padrão da performance do fundo, também denominado como a volatilidade do ativo.

3.4 Volatilidade e risco

A volatilidade é uma medida de risco que indica a magnitude das oscilações dos preços dos ativos em torno da média Markowitz (1952). Quanto maior a volatilidade, maior o risco. O método de Markowitz permite ao investidor escolher uma carteira que minimize o risco para um dado nível de retorno ou maximize o retorno para um dado nível de risco. No contexto dos Fundos de investimento imobiliários (FIIs) ou *Real estate investment trust* (REITs), a volatilidade e o risco são fatores importantes a serem considerados na seleção e gestão da carteira desses ativos. A volatilidade pode ser influenciada por diversos fatores, como as condições econômicas, políticas e sociais do país ou região onde os imóveis estão localizados.

O risco também é um fator importante na seleção e gestão da carteira de FIIs ou REITs e pode ser influenciado por diversos fatores, como a qualidade dos imóveis na carteira do fundo, a qualidade do inquilino e a qualidade do gestor do fundo. Liow e Song (2019) discutem a importância da volatilidade e do risco na tomada de decisão de investimentos em imóveis comerciais. Além disso, Liow e Zhu (2022) analisam a conexão entre a volatilidade dos preços dos REITs em diferentes países. Uma alta volatilidade indica que o preço do ativo está sujeito a grandes flutuações, enquanto uma baixa volatilidade indica que o preço do ativo é mais estável.

3.5 *Drawdown*

O *drawdown* é um conceito que expressa a maior queda do valor de um ativo ou de uma carteira em relação à sua cotação máxima em um determinado período (CHEKHLOV, URYASEV, ZABARANKIN, 2005). Ele é utilizado para medir o risco e a volatilidade de um investimento, bem como para comparar o desempenho de diferentes ativos ou estratégias. Esse conceito auxilia na gestão de portfólios, pois permite ao investidor avaliar o histórico de oscilação dos ativos e definir o nível de risco aceitável. Os mesmos autores propõem uma nova família de medidas de risco chamada *Drawdown Condicional* (CDD), que generaliza o *drawdown* para

um caso multi-cenário. Ademais, os autores desenvolvem técnicas para o cálculo do CDD e a solução de problemas de alocação de ativos com o CDD como medida de risco. Um alto *drawdown* indica uma grande queda em valor, enquanto um baixo *drawdown* indica quedas menores. Diante da pesquisa e suas implicações práticas, esse método pode ser utilizado para otimizar a alocação de ativos, buscando minimizar a perda máxima que pode ser experimentada por um investidor.

3.6 Algoritmos utilizados da FinRL

Na presente pesquisa, os objetos de análise principais foram os cinco algoritmos de aprendizado por reforço: *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 escolhidos por representarem padrões disponíveis na coleção FinRL, um arcabouço criado com o propósito de estudo e aplicação no setor financeiro do aprendizado por reforço (RL). A escolha desses métodos é baseada na representatividade de cada um em relação às principais técnicas e progressos no âmbito do aprendizado por reforço. Ademais, a presença prévia desses métodos na coleção FinRL torna mais fácil a comparação e avaliação entre eles.

3.6.1 *Advantage Actor-Critic* (A2C)

O método *Advantage Actor-Critic* (A2C) emprega uma função de vantagem para diminuir a flutuação do gradiente de políticas. Em vez de somente estimar a função de valor, a rede crítica avalia a função de benefício. (Mnih et al., 2016). Esta técnica é um aprimoramento do *Actor-Critic* (Rosenstein et al., 2004), que une um esquema de política - o ator - com uma função de valor - o crítico - para aprender uma política ideal. Sendo assim, o A2C é capaz de compreender espaços de ação dinâmicos e gerar observações de alta dimensão, assim como absorver padrões de políticas que sejam estáveis, em ambientes junto a múltiplos agentes (Wang et al., 2016).

Segundo Dang (2020), a fórmula para o A2C é definida por:

$\nabla J\theta\ (\theta)$ = E[Õ $T\ t$=1 $\nabla\theta$ log $\pi\theta\ (at\ |st)A(st\ ,\ at)$]

Deng et al. (2016) interpreta que $\pi\theta(at|st)$ é a rede de políticas, $A(st,\ at)$ é a função de vantagem que pode ser escrita como:

$A(st,\ at) = Q(st,\ at)\ -V\ (st)$

ou sob a perspectiva de Dhariwal et al. (2017):

$A(st,\ at) = r(st,\ at,\ st{+}1) + \gamma V\ (st{+}1)\ -V\ (st)$.

3.6.2 *Deep Deterministic Policy Gradient* (DDPG)

Já o *Deep Deterministic Policy Gradient* (DDPG) (Lillicrap et al., 2015) é empregado para promover o retorno máximo do investimento, o algoritmo reúne *frameworks* de *Q-learning* (LEARNING; SUTTON; BARTO, 1998) e *policy gradient* (SUTTON et al., 2000) em seu método e inclui redes neurais como mapeamento. Portanto, o DDPG assimila padrões diretamente das observações por meio do gradiente político, de modo a estruturar estados para ações de forma determinística e para melhor se adequar ao ambiente 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:

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

Sob a perspectiva de (FANG; LIU; YANG, 2019), a rede crítica é atualizada minimizando a função de perda $L(\theta Q)$, que é a diferença esperada entre as saídas da rede crítica alvo Q' e da rede crítica Q, ou seja,

$L(\theta Q)$ = E[$(yi - Q(st, at|\theta Q))$^2].

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).

3.6.3 *Proximal Policy Optmization* (PPO)

O *Proximal Policy Optmization* (PPO), conforme (Schulman et al., 2017), é utilizada para gerenciar a atualização do gradiente de política e assegurar que a política atualizada não possua distinções notáveis da anterior. O PPO procura simplificar o propósito da Otimização de Política de Região de Confiança (TRPO) ao adicionar um termo de corte na função objetivo (SCHULMAN et al., 2015) e (SCHULMAN et al., 2017). Sendo assim, supondo que a proporção de probabilidade entre as políticas antigas e novas seja expressa por:

$rt\ (\theta) = \pi\theta\ (at\ |st)\ |\ \pi\theta old\ (at\ |st)$.

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

J CLIP(θ) =E^t [min($rt\ (\theta)A$^$(st\ , at)$, clip($rt\ (\theta)$, 1 − ϵ, 1 + ϵ)A^$(st\ , at)$)].

Onde $rt\ (\theta)A$^(st, at) é o objetivo normal do gradiente de política e A^(st, at) é a função de vantagem estimada. A função $clip(rt\ (\theta)$, 1− ϵ, 1 + ϵ) corta a proporção $rt\ (\theta)$ 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.

3.6.4 *Soft Actor-Critic* (SAC)

O *Soft Actor-Critic* (SAC) é um algoritmo *off-policy* de aprendizado por reforço que combina o ator-crítico com a otimização da entropia máxima, buscando uma exploração mais eficiente e um equilíbrio entre exploração e explotação (Haarnoja et al., 2018). O SAC utiliza entropia para incentivar a exploração de informações, com uma visão distribucional do objetivo (Ma et al., 2020). O SAC tem sido aplicado com sucesso em diversas áreas, como robótica e simulações de controle, mostrando sua capacidade de resolver problemas complexos e de alta dimensão (Han & Sung, 2021). Sendo assim, o êxito do SAC em diversas áreas ressalta sua habilidade em lidar com problemas complexos e de alta dimensão, podendo interpretar o alto volume de negociações de FIIs, na presente pesquisa.

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

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

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.

3.6.5 *Twin Delayed Deep Deterministic Policy Gradient* (TD3)

O algoritmo de aprendizado por reforço proposto por Fujimoto, Hoof e Meger (2018) é conhecido como *Twin Delayed Deep Deterministic Policy Gradient* (TD3). O objetivo do TD3

é aumentar a estabilidade e o desempenho do *Deep Deterministic Policy Gradient* (DDPG) (Lillicrap et al., 2015). O TD3 faz diversas mudanças importantes, que incluem a atualização atrasada do ator, a utilização de duas redes críticas e a adição de ruído de ação direcionada. Isso proporciona não só um aumento da força do aprendizado de controle contínuo, como também na estabilidade.

O uso de duas redes cruciais para estimar a função Q é uma das principais inovações do TD3. Isso é feito para resolver o problema de superestimação de valor dos algoritmos baseados em *Q-learning*. O TD3 calcula o valor-alvo utilizando o mínimo dos valores estimados pelas duas redes críticas, limitando assim a superestimação. Portanto, para melhorar ainda mais o desempenho, o TD3 introduz um atraso na atualização do ator e da rede crítica-alvo, com o desenvolvimento das equações realizadas na obra de Fujimoto, Hoof e Meger (2018).

4. Análise dos resultados

4.1 Gráfico de *Advantage Actor-Critic* (A2C)

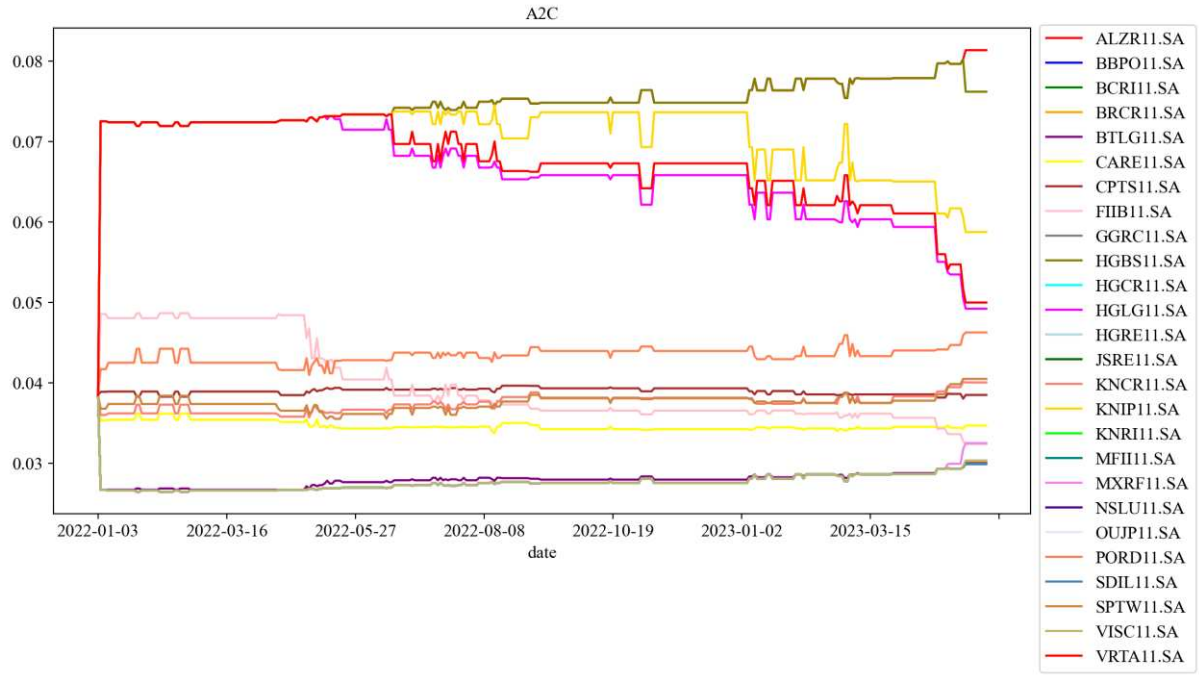

FIGURA 1 - Gráfico do desempenho do *Advantage Actor-Critic* (A2C). Fonte: Elaboração própria.

Com a composição da carteira estabelecida pelos 26 FIIs anteriormente citados e com 1% de custos de transação – assim como na elaboração dos demais gráficos dos algoritmos -, é possível notar, a partir da Figura 1, que houve movimentações mais notáveis de compra e venda em um grupo de ativos da carteira (HGBS11, KNIP11, HGLG11 e VRTA11). Considerando que, dentre estes, o fundo de *shoppings* (HGBS11) teve uma maior participação na carteira, assim como os fundos de títulos e valores mobiliários, setores industriais e logística obtiveram uma relativa queda na carteira entre, aproximadamente Junho de 2022 e Abril de 2023, algumas inferências surgem a partir deste contexto. Em primeiro lugar, sob o aspecto econômico, o aumento da participação do HGBS11 a partir de Junho de 2022 pode ter ocorrido pelo fato de que em Abril de 2022 os valores totais a receber deste fundo (aluguel, venda e outros) representavam cerca de R$ 15,3 milhões, passando para R$ 34,1 milhões no mês seguinte, margem de recebíveis nunca alcançada nos últimos 5 anos.

Segundamente, sob o aspecto político-econômico, é possível que a mudança na composição da carteira tenha sido influenciada, também, por eventos como a reunião do COPOM e a reunião do FED (próximas a Abril de 2022). A retórica mais amena adotada pelo COPOM e a

sinalização de uma pausa na elevação da taxa de juros pelo FED podem ter impactado positivamente alguns setores do mercado financeiro, levando a uma maior participação do fundo de shoppings (HGBS11) na carteira. Além disso, a queda nos preços das commodities pode ter afetado os setores industriais e logística, levando a uma queda na participação desses fundos na carteira, já que os valores a receber do fundo HGLG11 reduziram de uma média de R$ 200 milhões para R$ 124 milhões entre Maio e Setembro de 2022.

4.2 Gráfico de *Proximal Policy Optmization* (PPO)

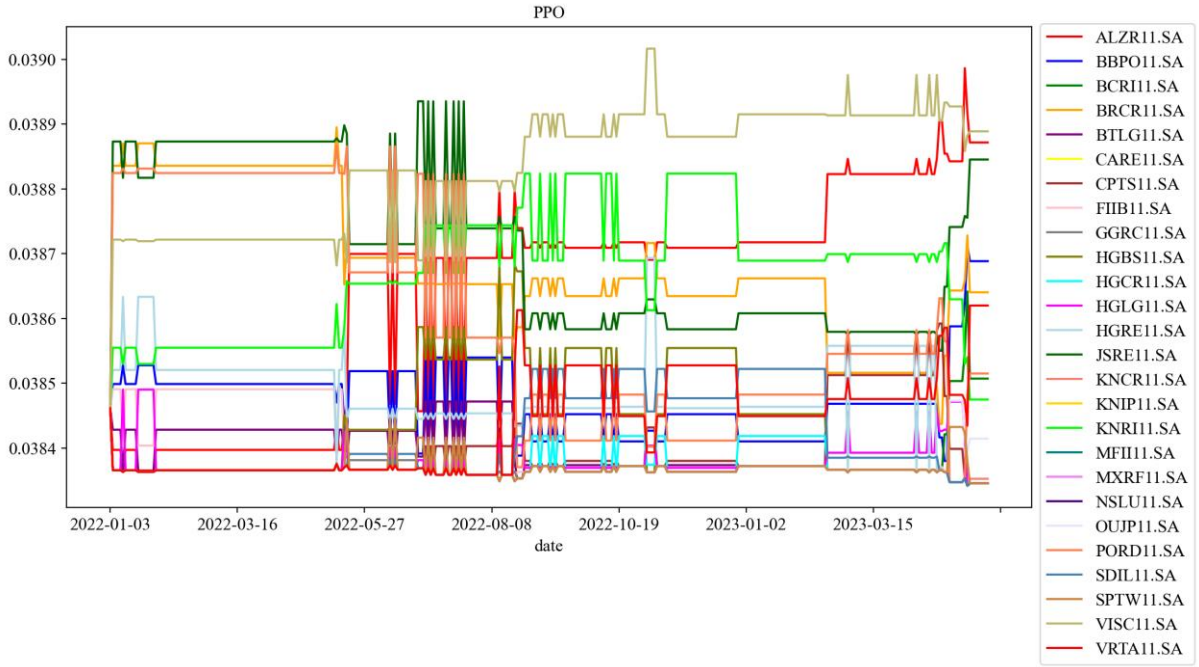

FIGURA 2 - Gráfico do desempenho do *Proximal Policy Optmization* (PPO). Fonte: Elaboração própria

Já em relação ao gráfico do algoritmo *Proximal Policy Optmization* (PPO), é possível observar que todo o período de pesquisa obteve mais movimentações em operações de compra e venda no geral, do que comparado com o A2C. Os dois períodos com mais oscilações notáveis, nesse contexto, são I) entre aproximadamente Junho e Julho de 2022 e II) entre Setembro e Outubro de 2022. O primeiro momento coincide com um período de oscilações na carteira também visto no A2C.

Tanto no primeiro período de análise, quanto no segundo, são observadas maiores alterações em fundos compostos por fundos híbridos, lajes corporativas, títulos e valores mobiliários em papel, *shoppings* e industriais e logísticos. Sendo assim, é possível notar que os tipos e segmentos Anbima dos fundos são os mesmos, apesar de serem ativos diferentes observados nestes dois momentos. Possíveis inferências acerca de tais variações é que estas podem ter sido influenciadas, também, pelas resoluções das reuniões do COPOM e do FED, no período em comum, no que diz respeito aos 2 algoritmos por hora observados, já que os tipos de fundos afetados foram os mesmos.

Sendo assim, tanto no momento I, quanto no momento II, a linguagem mais moderada que o COPOM empregou, mantendo a taxa SELIC em 13,75%, e a indicação do FED de que o aumento das taxas de juros iria estagnar, com o FED acrescendo à taxa de juros dos EUA 0,25%, podem ter gerado otimismo no mercado financeiro, levando a uma maior demanda por fundos de *shoppings* e títulos e valores mobiliários. Além disso, a alta inflação de março, medida pelo IPCA, foi de 1,62%, (a maior para esse mês desde 1994, além do o IPCA-15 de abril de 1,73%, abaixo das expectativas do mercado) o que pode ter afetado o comportamento dos fundos de lajes corporativas e industriais e logísticos, que podem ser mais sensíveis às variações de preços.

O ciclo de aperto monetário nos Estados Unidos também pode ter influenciado o comportamento desses fundos, tornando a taxa de juros brasileira mais competitiva para atração de investimentos.

Os gráficos gerados pelas estratégias dos algoritmos de *Deep Deterministic Policy Gradient* (DDPG), *Soft Actor-Critic* (SAC) e *Twin Delayed Deep Deterministic Policy Gradient* (TD3) não mostraram comportamentos que pudessem gerar inferências distintas das apresentadas pelos dois gráficos analisados e apresentaram um número menor de operações, em comparação com os métodos de A2C e PPO.

4.3 Gráfico de Retorno Acumulado

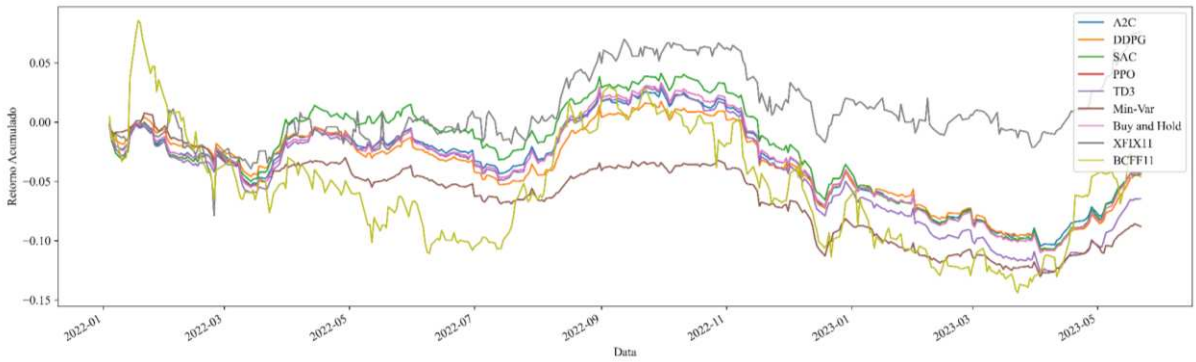

FIGURA 3 - Gráfico do Retorno Acumulado – Estratégia conjunta. Fonte: Elaboração própria

Em relação aos resultados obtidos com o Gráfico de Retorno Acumulado, pontos relevantes que merecem destaque são a proximidade do desempenho das carteiras elaboradas pelas estratégias de A2C, DDPG, PPO, SAC e *Buy and Hold*, sendo que dentre os algoritmos da FinRL utilizados, o SAC se sobressaiu positivamente, principalmente, no período entre Abril e Dezembro de 2022, em relação aos demais métodos. Uma possível explicação para isso é que o método de *Soft Actor-Critic* (SAC) foi responsável por focar a maior parte dos investimentos da carteira em VRTA11 e, deste modo, o ativo pode ter tido um destaque diante de FIIs dos setores híbridos, lajes corporativas, shoppings e industriais e logísticos pois pode ter se beneficiado de fatores econômicos, como a alta inflação, a valorização do real e a queda das taxas das NTNBs. Além disso, a perspectiva de um superciclo das commodities e a abertura de uma janela de oportunidade para estabilizar a economia brasileira de forma promissora podem ter contribuído para o desempenho do ativo.

Além disso, outro ponto de destaque é que a carteira dos algoritmos do aprendizado por reforço acompanhou o rendimento do XIFIX11 nos primeiros meses do período de análise, mas após Julho de 2022, o índice superou as carteiras elaboradas, com uma vantagem alta em relação a estes e a todas as outras estratégias do estudo realizado. Provavelmente, uma explicação para a diferença de desempenho observada é que como o IFIX compõe 111 fundos, a diversificação é maior do que cada uma das carteiras dos algoritmos individuais (com os mesmos 26 FIIs cada uma), como sustentado por Markowitz (1952), a diversificação permite ao investidor combinar ativos com baixa correlação entre si, o que a torna uma estratégia importante para reduzir o risco total da carteira, sendo uma estratégia importante na teoria de carteiras porque reduz o risco total da carteira sem reduzir o retorno esperado.

No entanto, a estratégia de mínima variância e de *Buy and Hold* não conseguiram superar os algoritmos, obtendo um desempenho bem abaixo e igual, respectivamente, as carteiras elaboradas, isso pode ter ocorrido pois uma carteira de mínima variância busca minimizar o risco, enquanto os algoritmos como DDPG, SAC, TD3, A2C e PPO, possuem ter diferentes

objetivos e estratégias de investimento, visando em sua maioria entender padrões e maximizar o retorno. Portanto, é possível que as carteiras montadas por esses algoritmos tenham um perfil de risco diferente da carteira de mínima variância e, consequentemente, apresentem um desempenho diferente, que neste caso, foi superior.

5. Conclusão

Neste estudo, analisou-se o desempenho de diferentes algoritmos de aprendizado por reforço na elaboração de carteiras de investimento compostas por Fundos de Investimento Imobiliário (FIIs). Os resultados mostraram que os algoritmos A2C, DDPG, PPO, SAC, TD3 e *Buy and Hold* apresentaram desempenhos semelhantes entre si e superiores a estratégias de mínima variância. Além disso, observou-se que eventos econômicos e políticos, como as reuniões do COPOM e do FED, bem como variações nos preços das commodities e na inflação, podem ter influenciado o comportamento dos fundos na carteira.

É importante notar que este estudo segue a Hipótese de Mercado Adaptativo, em vez da Hipótese de Mercado Eficiente. Isso significa que, embora a HME afirme que os preços futuros não podem ser previstos, acredita-se que os padrões podem ser compreendidos com o uso de algoritmos e comprovados em pesquisas futuras. Uma das limitações deste estudo é que ele se concentrou em um período específico e em um conjunto limitado de FIIs. Além disso, outros fatores que podem ter influenciado o desempenho dos fundos na carteira não foram considerados.

Este estudo tem implicações importantes para investidores e gestores de carteiras que buscam utilizar algoritmos de aprendizado por reforço para otimizar suas estratégias de investimento. Os resultados sugerem que esses algoritmos podem ser eficazes na elaboração de carteiras com desempenho superior a estratégias tradicionais. É recomendado que futuros estudos ampliem o escopo desta pesquisa, considerando diferentes períodos de tempo e conjuntos mais amplos de FIIs. Além disso, seria interessante investigar o impacto de outros fatores econômicos e políticos no desempenho das carteiras elaboradas pelos algoritmos.

Em trabalhos futuros, espera-se que a pesquisa seja estendida por outros acadêmicos para incluir outros algoritmos de aprendizado por reforço e comparar seus desempenhos com outras estratégias de investimento. Também temos a expectativa da investigação do impacto de diferentes configurações dos algoritmos no desempenho das carteiras em trabalhos posteriores. Este estudo traz importantes contribuições para investidores e gestores de carteiras do mercado financeiro de maneira geral. Além disso, apresenta novas metodologias para a análise, o que pode auxiliar na tomada de decisões mais informadas e precisas. As implicações deste trabalho são significativas e podem ter um impacto positivo no setor financeiro como um todo. Sinceros agradecimentos aos nossos colaboradores e orientador pelo apoio a esta pesquisa.

Referências

ALDRIGHI, Dante Mendes; MILANEZ, Daniel Yabe. Finança comportamental e a hipótese dos mercados eficientes. **Revista de Economia Contemporânea**, v. 9, n. 1, 2005.

B3. **Boletim Mensal Fundos Imobiliários (FIIs)**. São Paulo, abril de 2023. Disponível em: https://www.b3.com.br/data/files/FC/D2/01/14/E020881064456178AC094EA8/Boletim_FII_-04M23.pdf. Acesso em: 9 jul. 2023.

B3. **Boletim Mercado Imobiliário**. São Paulo, junho de 2018. Disponível em: https://www.b3.com.br/data/files/87/04/49/B9/29AA4610C2E69A46AC094EA8/Boletim Mercado Imobiliário - 2018 06.pdf. Acesso em: 9 jul. 2023.

BRUEGGEMAN, William B.; FISHER, Jeffrey D. **Real estate finance and investments**. New York: McGraw-Hill Irwin, 2011.

CARHART, Mark M. et al. Mutual fund survivorship. **The review of financial studies**, v. 15, n. 5, p. 1439-1463, 2002.

CASTELLO BRANCO, Carlos Eduardo; MONTEIRO, Eliane de Mello Alves Rebouças. **Estudo sobre a indústria de fundos de investimentos imobiliários no Brasil**. 2003.

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

CHEKHLOV, Alexei; URYASEV, Stanislav; ZABARANKIN, Michael. Drawdown measure in portfolio optimization. **International Journal of Theoretical and Applied Finance**, v. 8, n. 01, p. 13-58, 2005.

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.

FAMA, Eugene F. et al. The adjustment of stock prices to new information. **International economic review**, v. 10, n. 1, p. 1-21, 1969.

FAMA, Eugene F.; FRENCH, Kenneth R. **Common risk factors in the returns on stocks and bonds. Journal of financial economics**, v. 33, n. 1, p. 3-56, 1993.

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.

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.

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

HARTZELL, David; HEKMAN, John S.; MILES, Mike E. **Real estate returns and inflation**. Real Estate Economics, v. 15, n. 1, p. 617-637, 1987

IÓRIO, Fabio Roberto et al. **Análise do desempenho de carteiras de fundos de investimento imobiliário negociados na BM&FBOVESPA entre 2011 e 2013**. 2014.

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

LINTNER, John. Security prices, risk, and maximal gains from diversification. **The journal of finance**, v. 20, n. 4, p. 587-615, 1965.

LINTNER, John. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets: A reply. **The review of economics and statistics**, p. 222-224, 1969.

LO, Andrew W. The adaptive markets hypothesis: **Market efficiency from an evolutionary perspective. Journal of Portfolio Management, Forthcoming**, 2004.

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

MARKOWITZ, H. M. **Portfolio Selection, the journal of finance**. 1952.

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

MOSSIN, Jan. Equilibrium in a capital asset market. **Econometrica: Journal of the econometric society**, p. 768-783, 1966.

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

NETO, Alexandre Assaf; 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.

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

ROSENSTEIN, Michael T. et al. Supervised actor-critic reinforcement learning. **Learning and Approximate Dynamic Programming: Scaling Up to the Real World**, p. 359-380, 2004.

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.; ALEXANDER, Gordon J.; BAILEY, Jeffery V. **Investment**. Prentice Hall Incorporated, 1999.

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.; MCALLESTER DAVID, A.; SINGH SATINDER, P. Mansour Yishay. Policy gradient methods for reinforcement learning with function approximation. **Advances in neural information processing systems**, p. 1057-1063, 2000.

TOBIN, James. Liquidity preference as behavior towards risk. **The review of economic studies**, v. 25, n. 2, p. 65-86, 1958.

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**. 2020. p. 1-8.

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.