

UNIVERSIDADE FEDERAL DE SANTA CATARINA

**APLICAÇÃO DE APRENDIZADO DE MÁQUINA PARA REBALANCEAMENTO DE
CARTEIRA DE INVESTIMENTOS**

Guilherme Cardoso de Oliveira

Florianópolis – SC

2026/ 1

Guilherme Cardoso de Oliveira

APLICAÇÃO DE APRENDIZADO DE MÁQUINA PARA REBALANCEAMENTO DE CARTEIRA DE INVESTIMENTOS

Trabalho de Conclusão de Curso submetido ao Curso de Graduação em Sistemas de Informação da Universidade Federal de Santa Catarina para a obtenção do Grau de Bacharel em Sistemas de Informação.

Orientador: André Wüst Zibetti.

Florianópolis, Santa Catarina – Brasil

2026

Guilherme Cardoso de Oliveira

APLICAÇÃO DE APRENDIZADO DE MÁQUINA PARA REBALANCEAMENTO DE CARTEIRA DE INVESTIMENTOS

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Sistemas de Informação, e foi aprovado em sua forma final pelo Curso de Graduação em Sistemas de Informação da Universidade Federal de Santa Catarina.

Florianópolis, Santa Catarina – Brasil, 25 de junho de 2026.

Banca Examinadora:

André Wüst Zibetti
Orientador

Elder Rizzon Santos
Avaliador

Alexandre Gonçalves Silva
Avaliador

RESUMO

O mercado financeiro brasileiro impõe desafios significativos à gestão de portfólio, dada sua elevada volatilidade e a influência de variáveis macroeconômicas, diante das quais os modelos tradicionais de alocação, como a Teoria Moderna do Portfólio, exibem limitada capacidade de adaptação dinâmica. Este trabalho investigou a viabilidade e a robustez de um modelo de Aprendizado por Reforço, baseado no algoritmo Proximal Policy Optimization (PPO), para o rebalanceamento dinâmico de uma carteira de cinco ETFs do mercado brasileiro (BOVA11 (renda variável), IMAB11 (títulos indexados ao IPCA), IRFM11 (títulos pré-fixados), XFIX11 (mercado imobiliário) e GOLD11 (ouro)), operando em um ambiente de simulação enriquecido com indicadores macroeconômicos (SELIC e IPCA) e com a fricção transacional realista da bolsa. Adotou-se como recompensa o log-retorno líquido de custos, e o desempenho do modelo foi avaliado sob protocolo multi-seed em três regimes macroeconômicos distintos: juros estáveis, choque pandêmico e aperto monetário, em confronto com as estratégias Buy & Hold, SELIC e Risk Parity por meio de testes t. Os resultados evidenciaram robustez condicional ao regime: o modelo superou o Buy & Hold sob juros estáveis, cedeu-lhe sob o choque agudo e foi dominado pelas estratégias de renda fixa sob o aperto monetário. A análise da composição da carteira revelou, ademais, que o modelo não adapta sua alocação ao regime macroeconômico, atuando como um alocador tático direcional cuja vantagem provém do reposicionamento de alta frequência, e não de uma estratégia condicionada ao contexto. Conclui-se pela viabilidade condicional da abordagem, delimitando-se as condições de competitividade do método e as direções para seu aprimoramento.

Palavras-chave: Aprendizado por Reforço, Otimização de Portfólio, PPO, Macroeconomia, ETFs.

ABSTRACT

The Brazilian financial market poses significant challenges to portfolio management, given its high volatility and the influence of macroeconomic variables, in the face of which traditional allocation models, such as Modern Portfolio Theory, exhibit limited capacity for dynamic adaptation. This work investigated the feasibility and robustness of a Reinforcement Learning model, based on the Proximal Policy Optimization (PPO) algorithm, for the dynamic rebalancing of a portfolio of five Brazilian ETFs — BOVA11 (equities), IMAB11 (inflation-linked bonds), IRFM11 (fixed-rate bonds), XFIX11 (real estate) and GOLD11 (gold) — operating in a simulation environment enriched with macroeconomic indicators (the SELIC policy rate and the IPCA inflation index) and with realistic exchange transaction friction. The cost-adjusted log-return was adopted as the reward function, and the model's performance was assessed under a multi-seed protocol across three distinct macroeconomic regimes: stable interest rates, pandemic shock and monetary tightening, in comparison with the Buy & Hold, SELIC and Risk Parity strategies by means of t-tests. The results evidenced regime-conditional robustness: the model outperformed Buy & Hold under stable interest rates, yielded to it under the acute shock, and was dominated by fixed-income strategies under monetary tightening. Analysis of the portfolio composition further revealed that the model does not adapt its allocation to the macroeconomic regime, acting instead as a directional tactical allocator whose advantage stems from high-frequency repositioning rather than from a context-conditioned strategy. It is concluded that the approach is conditionally feasible, with the conditions of the method's competitiveness and the directions for its improvement being delimited.

Keywords: Reinforcement Learning, Portfolio Optimization, PPO, Macroeconomics, ETFs.

LISTA DE ILUSTRAÇÕES

Figura 1 – Fluxo Metodológico da Pesquisa	31
Figura 2 - Métricas internas de treinamento do Lote 1: clip_fraction	49
Figura 3 - Métricas internas de treinamento do Lote 1: approx_kl.....	50
Figura 4 - Métricas internas de treinamento do Lote 1: explained_variance	50
Figura 5 - Histograma dos retornos das execuções - cenário principal (2023–2024)	73
Figura 6 - Evolução da dispersão entre execuções - cenário principal (2023–2024)	74
Figura 7 - Composição média da carteira do modelo - cenário principal (2023–2024)	75
Figura 8 - Composição representativa de uma carteira do modelo - cenário principal (2023 - 2024).....	75
Figura 9 - Curva de capital multi-seed e estratégias de referência - cenário principal (2023–2024).....	76
Figura 10 - Mapa de calor dos retornos mensais - cenário principal (2023–2024)....	77
Figura 11- Curvas de drawdown das estratégias - cenário principal (2023–2024)....	78
Figura 12 - Curva de capital multi-seed e estratégias de referência - cenário COVID- 19 (2020).....	81
Figura 13 - Histograma dos retornos das execuções - cenário COVID-19 (2020)	82
Figura 14 - Curva de capital multi-seed e estratégias de referência - cenário de expansão da SELIC (2021–2022)	84
Figura 15 - Retorno por execução versus número de rebalanceamentos (bimodalidade) - cenário de expansão da SELIC (2021-2022)	85
Figura 16 - Distribuição dos retornos do modelo nos três regimes (diagramas de caixa).....	88
Figura 17 - Divisão temporal dos dados em treino, validação e teste por cenário	88

LISTA DE TABELAS

Tabela 1 – Comparação dos tipos de Aprendizado por Reforço	17
Tabela 2 – Posicionamento comparativo dos trabalhos relacionados.....	29
Tabela 3 – Construção de Séries Proxy e Validação por Correlação.....	37
Tabela 4 - Configuração final do modelo.....	60
Tabela 5 - Estatísticas descritivas do cenário principal (2023–2024).....	69
Tabela 6 – Análise inferencial do desempenho frente às estratégias de referência (cenário principal, 2023 - 2024).....	71
Tabela 7 - Estatísticas descritivas do cenário de estresse pandêmico (COVID-19, 2020)	79
Tabela 8 - Estatísticas descritivas do cenário de aperto monetário (expansão da SELIC, 2021-2022).....	83
Tabela 9 - Síntese comparativa do desempenho entre regimes macroeconômicos .	86

SUMÁRIO

1	INTRODUÇÃO	10
1.1	PROBLEMA E JUSTIFICATIVA	11
1.2	OBJETIVO GERAL.....	13
1.3	OBJETIVOS ESPECÍFICOS	13
2	FUNDAMENTAÇÃO TEÓRICA.....	14
2.1	APRENDIZADO DE MÁQUINA	14
2.2	APRENDIZADO POR REFORÇO (RL)	15
2.3	CLASSIFICAÇÃO DE ALGORITMOS DE APRENDIZADO POR REFORÇO ..	16
2.3.1	Model-Based vs Model-Free	17
2.3.2	Value-Based, Policy-Based e Actor-Critic	18
2.3.3	O Algoritmo Proximal Policy Optimization (PPO)	20
2.4	REBALANCEAMENTO DE PORTFÓLIO	21
2.4.1	Custos do rebalanceamento	22
2.5	CONCEITOS ECONÔMICOS E MÉTRICAS FINANCEIRAS	23
2.5.1	Exchange Traded Funds (ETFs)	23
2.5.2	Variáveis Macroeconômicas	23
2.5.3	Métricas de Avaliação de Desempenho	24
2.6	FUNÇÃO DE RECOMPENSA.....	26
2.7	TRABALHOS RELACIONADOS	26
3	METODOLOGIA	30
3.1	VISÃO GERAL DO FLUXO METODOLÓGICO	30
3.2	DEFINIÇÃO DO UNIVERSO DE ATIVOS	33
3.3	COLETA E TRATAMENTO DE DADOS	34
3.3.1	Construção de séries Proxy e validação por Correlação	35
3.4	CONSTRUÇÃO DO AMBIENTE DE SIMULAÇÃO	40
3.4.1	Espaço de Estados	41
3.4.2	Espaço de Ações e Restrição de Concentração	42
3.4.3	Função de Recompensa	44
3.4.4	Dinâmica de Transação e Custo de Rebalanceamento	45

3.5	PROCESSO ITERATIVO DE REFINAMENTO DO AGENTE	47
3.5.1	Configuração Inicial e Diagnóstico de Instabilidade	48
3.5.2	Ajuste de Hiperparâmetros e Emergência de Hiperatividade	52
3.5.3	Investigação de Função de Recompensa Penalizada pela Taxa Livre de Risco	54
3.5.4	Restrição Estrutural de Concentração e o Surgimento do Atrito Operacional	56
3.6	CONFIGURAÇÃO FINAL E PROTOCOLO MULTI-SEED	60
3.6.1	Configuração final	60
3.6.2	Controle de Estocasticidade e Protocolo Multi-Seed	61
3.6.3	Apresentação Estatística dos Resultados	62
3.7	PERÍODOS DE AVALIAÇÃO E CENÁRIOS DE TESTE	62
3.8	ESTRATÉGIA DE AVALIAÇÃO E BASELINES	65
4	RESULTADOS	68
4.1	VISÃO GERAL	68
4.2	CENÁRIO PRINCIPAL	69
4.2.1	Estatísticas Descritivas Agregadas	69
4.2.2	Análise Inferencial vs. Baselines	71
4.2.3	Distribuição dos Desempenhos	72
4.2.4	Trajetória e Comportamento Operacional	75
4.2.5	Síntese do Cenário Principal	78
4.3	CENÁRIO DE ESTRESSE PANDÊMICO: COVID-19 (2020)	79
4.4	CENÁRIO DE EXPANSÃO DA SELIC (2021 - 2022)	82
4.5	ANÁLISE COMPARATIVA ENTRE REGIMES	86
4.6	DISCUSSÃO DOS RESULTADOS	90
5	CONCLUSÃO	94
6	REFERÊNCIAS BIBLIOGRÁFICAS	96

1 INTRODUÇÃO

O mercado financeiro é um ecossistema caracterizado por sua dinâmica não-linear e alta volatilidade, onde a incerteza é a única constante. Nesse ambiente, a gestão de ativos ultrapassa a simples seleção de investimentos rentáveis, trata-se de um exercício contínuo de gerenciamento de riscos e adaptação a cenários adversos com grande implicação de fatores externos. Como alertou Warren Buffett, "O risco vem de não saber o que você está fazendo" (Buffett, citado por Hagstrom, 2005). Essa máxima ressalta que, em um sistema complexo, a vulnerabilidade não reside apenas na flutuação dos preços, mas na incapacidade dos modelos de gestão de interpretarem corretamente o regime de mercado vigente e reagirem de forma adequada.

Historicamente, a Teoria Moderna do Portfólio (MPT), proposta por Markowitz, serviu como base para a alocação e gerenciamento de ativos. Embora de extrema importância teórica para a diversificação, quando aplicada à realidade de mercados emergentes e recessivos, verifica-se uma carência fundamental de adaptação dinâmica. Os modelos clássicos dependem de parâmetros estáticos, como vetores de retornos esperados e matrizes de covariância baseados exclusivamente em dados passados, tornando-os altamente sensíveis a erros de estimação e pouco resilientes diante de choques e mudanças abruptas no cenário econômico (Uppaladinni et al., 2024).

Nesse cenário de complexidade, a manutenção da performance de longo prazo depende intrinsecamente do rebalanceamento de portfólio. Este processo consiste no realinhamento periódico das ponderações dos ativos para restaurar a fronteira eficiente e evitar exposições sistêmicas não planejadas. No entanto, as estratégias tradicionais de rebalanceamento enfrentam o dilema prático entre a necessidade de equacionar os gastos transacionais de execução contra as potenciais perdas decorrentes de uma carteira com exposição subótima. Conforme elucidado por Masters (2003), o gestor lida com um *trade-off* constante entre aceitar o desvio estratégico da carteira ou incorrer em elevados custos de transação para readequar os pesos. Estratégias determinísticas muitas vezes ignoram essa fricção de mercado e o estado macroeconômico vigente, resultando em execuções ineficientes que corroem a rentabilidade real do portfólio.

É para superar essa rigidez que o Aprendizado por Reforço (Reinforcement Learning) se apresenta como uma evolução necessária. Ao modelar o rebalanceamento como um problema de decisão sequencial (MDP), a técnica permite que o investidor utilize uma política de rebalanceamento dinâmica, que se adapta às condições do ambiente, ao invés de regras engessadas. A literatura recente valida este caminho, Ferreira (2025) aplicou RL para o rebalanceamento de cripto ativos, enquanto Almahdi e Yang (2017) exploraram arquiteturas de risco para mitigar perdas severas. Contudo, a aplicação de algoritmos densos no mercado brasileiro demanda a superação do impasse supracitado, onde abordagens focadas na maximização desenfreada do retorno frequentemente colapsam ao enfrentarem as taxas de corretagem e a liquidez real das bolsas de valores.

Dessa forma, este trabalho se posiciona na intersecção da ciência da computação e das finanças quantitativas, propondo um modelo de rebalanceamento inteligente fundamentado no algoritmo Proximal Policy Optimization (PPO) aplicado a uma cesta diversificada de fundos de índice no mercado brasileiro. A abordagem concentra sua inovação no design estrutural do ambiente de simulação, o qual integra um espaço de estados enriquecido por variáveis macroeconômicas, cuja relevância direta na precificação de ativos é respaldada pela literatura (Bartkus; Paleviciene, 2013). O objetivo central consiste em investigar a viabilidade e a robustez de um modelo assim constituído, avaliando seu desempenho frente a estratégias de referência ao longo de regimes macroeconômicos distintos, e em caracterizar o comportamento de alocação que dele emerge sob a fricção transacional do mercado brasileiro.

1.1 PROBLEMA E JUSTIFICATIVA

A prática da gestão de portfólios no mercado financeiro contemporâneo esbarra em limitações estruturais intrínsecas às abordagens estáticas, cuja ineficiência diante de choques macroeconômicos e quebras de regime é amplamente documentada na literatura (Guidolin; Timmermann, 2007). De um lado, modelos de otimização clássicos, fundamentados na Teoria Moderna do Portfólio (Markowitz, 1952), estabeleceram as bases matemáticas inegáveis para a diversificação, mas exibem severa obsolescência temporal na gestão ativa. Conforme demonstrado pelo estudo

empírico de DeMiguel, Garlappi e Uppal (2009), a dependência desses modelos em relação a retornos médios e matrizes de covariância históricas gera erros de estimação sistêmicos que degradam a performance fora da amostra (out-of-sample), tornando-os inflexíveis e incapazes de capturar a natureza não estacionária dos ativos. Por outro lado, a gestão puramente discricionária demonstra profunda vulnerabilidade analítica sob estresse sistêmico. Análises recentes sobre carteiras institucionais (PUC Minas, 2023) evidenciam que a alocação de pesos baseada no julgamento humano tende a agravar perdas durante crises econômicas, subperformando significativamente em relação a métodos de controle algorítmico de risco. Essa dupla deficiência operacional reforça a urgência de substituir heurísticas subjetivas e matrizes estáticas por sistemas de decisão dinâmica, mitigando o viés comportamental e garantindo uma readequação com a velocidade computacional que a volatilidade real exige.

Simultaneamente, tentativas de modelar o mercado brasileiro via estatística tradicional mostram-se limitadas. Theodoro (2018), ao aplicar modelos de Vetores Autorregressivos (VAR), concluiu que a relação linear entre variáveis macroeconômicas (SELIC, IPCA) e o mercado acionário (IBOVESPA) é estatisticamente inconclusiva sob uma ótica estritamente linear. Essa falha não indica a ausência de relação, mas sugere que a interação entre macroeconomia e a precificação dos ativos é intrinsecamente complexa e não-linear. Ademais, aplicações basilares de Aprendizado por Reforço em ativos digitais (Ferreira, 2025) evidenciam que abordagens desprovidas de um estado macroeconômico robusto, e que ignoram as fricções transacionais de mercado, tendem a falhar ou desenvolver políticas hiperativas e financeiramente insustentáveis.

Portanto, a problemática central que norteia esta pesquisa reside na urgência por um modelo adaptativo capaz de capturar a não-linearidade do cenário macroeconômico e, simultaneamente, viabilizar execuções lucrativas na presença da fricção transacional inerente à bolsa de valores. O Aprendizado por Reforço justifica-se como a abordagem metodológica ideal por tratar a alocação de capital como um Processo de Decisão de Markov (MDP), permitindo a indução de políticas verdadeiramente dinâmicas a partir dos dados brutos do ambiente. A escolha técnica do algoritmo Proximal Policy Optimization (PPO) baseia-se em sua estabilidade matemática comprovada para a convergência de políticas em espaços de ações

contínuos (Schulman et al., 2017), superando as limitações de discretização inerentes a arquiteturas puramente baseadas em valor. Assim, este trabalho busca investigar em que medida um modelo de RL oferece desempenho mais robusto que os modelos estáticos diante das flutuações e quebras de ciclos de juros no mercado brasileiro.

1.2 OBJETIVO GERAL

Desenvolver e avaliar um Modelo de Aprendizado por Reforço (RL), utilizando o algoritmo Proximal Policy Optimization (PPO), para o rebalanceamento dinâmico de um portfólio composto por Exchange Traded Funds (ETFs) de índice no mercado brasileiro e avaliá-lo com base em estratégias de referência.

1.3 OBJETIVOS ESPECÍFICOS

Para alcançar o objetivo geral, foram estabelecidos os seguintes objetivos específicos:

- Realizar uma revisão bibliográfica dos métodos de otimização de portfólio e na utilização de técnicas de machine learning no mercado financeiro;
- Constituir uma base de dados íntegra e alinhada temporalmente, integrando cotações de ETFs selecionados com indicadores macroeconômicos (SELIC e IPCA) provenientes de fontes oficiais;
- Construir um ambiente de simulação do mercado, capaz de processar um estado realista com dados de mercado e variáveis macroeconômicos;
- Implementar e treinar um modelo de *Reinforcement Learning* empregando a arquitetura algorítmica do PPO;
- Analisar a robustez e a capacidade de adaptação do modelo treinado em diferentes períodos de mercado, como períodos de crise e de alta volatilidade, avaliando a capacidade de generalização do modelo treinado;
- Comparar quantitativamente o desempenho do modelo treinado com estratégias passivas de referência (*benchmarks*), como o *Buy & Hold*.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os principais conceitos que servem de alicerce para o trabalho. O percurso argumentativo parte da definição do Aprendizado de Máquina e de seus paradigmas, adentrando progressivamente na especificidade do Aprendizado por Reforço e na formalização matemática do Processo de Decisão de Markov. Na sequência, justifica-se a escolha do algoritmo Proximal Policy Optimization (PPO) pela análise taxonômica das famílias algorítmicas do campo. A seção se encerra com a fundamentação das práticas de rebalanceamento de portfólio, dos custos de fricção de mercado e das métricas financeiras empregadas na avaliação dos resultados.

2.1 APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina (Machine Learning - ML) constitui um subcampo da Inteligência Artificial centrado no desenvolvimento de algoritmos capazes de extrair padrões de dados e aprimorar seu desempenho com a experiência, sem a necessidade de serem explicitamente programados para cada tarefa particular (MITCHELL, 1997). A distinção central em relação à programação clássica reside precisamente nesse mecanismo: enquanto o paradigma convencional exige que o programador enuncie regras lógicas a priori, os algoritmos de ML induzem essas regras a partir de grandes volumes de dados históricos, conferindo ao modelo uma capacidade de generalização para instâncias inéditas, característica essencial em domínios como o mercado financeiro, onde o passado jamais se repete com exatidão.

A literatura classifica os algoritmos de ML em três grandes paradigmas, diferenciados pela natureza do sinal de feedback utilizado durante o treinamento (MITCHELL, 1997). O Aprendizado Supervisionado opera sobre conjuntos de dados rotulados, nos quais cada entrada possui uma saída correta conhecida, e o objetivo do algoritmo consiste em aprender uma função de mapeamento que minimize o erro de predição sobre exemplos futuros (GOODFELLOW; BENGIO; COURVILLE, 2016). O Aprendizado Não Supervisionado, por sua vez, lida com dados desprovidos de rótulos, incumbindo o algoritmo de identificar estruturas latentes, agrupamentos ou representações compactas dos dados sem orientação externa (GOODFELLOW; BENGIO; COURVILLE, 2016). Contudo, para problemas que envolvem tomada de

decisão sequencial e adaptação contínua a ambientes dinâmicos, como o rebalanceamento de portfólios, ambos os paradigmas anteriores revelam limitações estruturais.

É nesse cenário que o Aprendizado por Reforço (Reinforcement Learning - RL) emerge como o paradigma mais adequado. Diferentemente do aprendizado supervisionado, o RL não depende de um conjunto estático de dados com respostas corretas predefinidas, mas sim da interação dinâmica entre um modelo decisor e um ambiente: o modelo observa o estado corrente, executa uma ação e recebe um sinal de recompensa como feedback de sua decisão (SUTTON; BARTO, 2018). Essa metodologia reproduz, com maior fidelidade, o comportamento de um gestor de investimentos, que deve realizar alocações contínuas visando maximizar o patrimônio no longo prazo, sem acesso prévio à trajetória correta das decisões.

2.2 APRENDIZADO POR REFORÇO (RL)

O Aprendizado por Reforço distingue-se fundamentalmente das demais abordagens por centrar-se no problema computacional de aprender a agir por meio da interação. Conforme a definição seminal de Sutton e Barto (2018), RL consiste em aprender o que fazer, como mapear situações em ações, de modo a maximizar um sinal numérico de recompensa cumulativa. O modelo não recebe instruções sobre qual ação tomar: ele deve descobrir, por tentativa e erro, quais ações produzem os maiores retornos, levando em conta não apenas a recompensa imediata, mas os efeitos das decisões presentes sobre os estados futuros.

A formalização matemática deste problema ocorre por meio do arcabouço dos Processos de Decisão de Markov (MDP). Um MDP é definido formalmente pela quintupla (S, A, T, R, γ) , em que: S é o espaço de estados que descreve as possíveis configurações do ambiente; A é o espaço de ações disponíveis ao modelo; $T: S \times A \times S \rightarrow [0,1]$ é a função de transição de estados, que representa a probabilidade de o ambiente passar do estado s para o estado s' após a execução da ação a ; $R: S \times A \rightarrow \mathbb{R}$ é a função de recompensa que quantifica a qualidade imediata de uma ação em um dado estado; e $\gamma \in [0,1)$ é o fator de desconto intertemporal, que determina o

quanto o modelo prioriza recompensas futuras em relação às imediatas (SUTTON; BARTO, 2018; PUTERMAN, 1994).

O objetivo do modelo consiste em encontrar uma política ótima $\pi^*: S \rightarrow A$, entendida como uma função que mapeia estados a ações, de forma a maximizar o retorno esperado acumulado e descontado:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

A cada instante de tempo t , o modelo observa o estado S_t do ambiente, seleciona uma ação A_t de acordo com sua política corrente, recebe a recompensa R_{t+1} enquanto o ambiente transita para o novo estado S_{t+1} . Em aplicações de Deep Reinforcement Learning, a política π é parametrizada por uma rede neural profunda, o que permite ao modelo lidar com espaços de estado de alta dimensionalidade sem incorrer na maldição da dimensionalidade que afeta métodos tabulares (BELLMAN, 1957; MNIH et al., 2015).

A correspondência entre os elementos abstratos do MDP e o problema concreto deste trabalho é direta. O estado S_t é composto por uma janela deslizante de 21 pregões de dados de mercado, contendo log-retornos diários, volatilidade e os índices de Sharpe e Sortino de janela móvel para cada um dos cinco ativos, acrescidos dos valores da taxa SELIC diária e do IPCA, que fornecem o contexto macroeconômico necessário para a identificação de quebras de regime. A ação A_t corresponde ao vetor de pesos da carteira, determinado pela aplicação de uma função softmax com restrição de concentração máxima de 40% por ativo sobre os logits contínuos emitidos pelo ator. A recompensa R_t é o log-retorno líquido do portfólio após a dedução de custos de transação, conforme detalhado na seção de metodologia.

2.3 CLASSIFICAÇÃO DE ALGORITMOS DE APRENDIZADO POR REFORÇO

A taxonomia dos algoritmos de RL estrutura-se em dois eixos complementares, cada um conduzindo a uma escolha de design diferente. O primeiro eixo distingue os algoritmos pelo conhecimento que o modelo possui sobre o ambiente. O segundo, pelo que o modelo aprende internamente, uma função de valor, uma política direta,

ou ambas simultaneamente. A compreensão dessas distinções é fundamental para justificar, com rigor, a escolha do algoritmo PPO para o problema de alocação de ativos.

2.3.1 Model-Based vs Model-Free

Na família *Model-Based*, o modelo constrói ou dispõe de um modelo explícito da dinâmica do ambiente, isto é, da função de transição $T(s, a, s')$ e da função de recompensa $R(s, a)$, utilizando esse modelo para planejar suas ações por simulação interna antes de interagir com o ambiente real. Essa abordagem é eficiente em termos de dados quando o modelo é preciso, mas exige que a dinâmica do domínio seja conhecida ou aprendível com baixo erro (SUTTON; BARTO, 2018).

O mercado financeiro, contudo, é intrinsecamente não-estacionário e governado por dinâmicas de transição desconhecidas e de difícil modelagem analítica. Preços de ativos dependem de expectativas futuras, comportamento de variáveis heterogêneas, informações assimétricas e eventos macroeconômicos, tornando qualquer modelo da transição de estados inerentemente sujeito a erros sistemáticos de especificação. Por essa razão, a família *Model-Free*, na qual o modelo aprende a política ótima diretamente a partir da experiência bruta, sem construir um modelo do ambiente, é amplamente adotada em aplicações financeiras de RL (MNIH et al., 2015; SCHULMAN et al., 2017). Esta é a abordagem empregada no presente trabalho.

Tabela 1 – Comparação dos tipos de Aprendizado por Reforço

Classificação	Model-Based	Model-Free
Conhecimento do ambiente	Constrói ou exige um modelo preditivo explícito da dinâmica de transição de estados e das recompensas do sistema.	Não modela a dinâmica internamente; aprende a política de ação ou a função de valor puramente pela experiência empírica.
Eficiência de dados	Permite o planejamento otimizado da política através de simulações internas computacionais antes da interação real.	Moderada a baixa. Exige um volume expressivamente maior de interações (tentativa e erro) para que a política convirja.

Requisito de precisão	Alto. Erros na especificação do modelo propagam-se cumulativamente	Robusto. Resistente a falhas de modelagem da dinâmica do sistema, pois adapta-se diretamente aos dados observados.
Aplicabilidade financeira	Limitado, devido à imprevisibilidade da dinâmica de mercado	Predominante (abordagem adotada). Apresenta resiliência ao ruído do mercado

Fonte: Elaboração própria (2026) base em Sutton e Barto (2018)

2.3.2 Value-Based, Policy-Based e Actor-Critic

Dentro da categoria Model-Free, os algoritmos diferenciam-se pelo objeto de aprendizado. Os métodos *Value-Based* aprendem uma função de valor que estima o retorno futuro esperado para cada par estado-ação, a chamada Q-função, denotada por $Q(s, a)$, e derivam a política de forma gulosa (*greedy*) em relação a esse valor estimado.

$$A_t = \operatorname{argmax}_a Q(S_t, a)$$

Nessa expressão, A_t representa a ação selecionada pelo modelo no instante t , escolhida como aquela que maximiza o valor estimado $Q(S_t, a)$ dentre todas as ações a disponíveis no estado corrente S_t . O modelo, portanto, estima o retorno futuro esperado de cada ação possível e seleciona deterministicamente a de maior valor, sem considerar alternativas de menor valor imediato que porventura conduzissem a estados mais vantajosos no longo prazo. Essa regra de decisão localmente ótima é denominada política gulosa (*greedy*) em relação à função Q (SUTTON; BARTO, 2018).

O algoritmo Deep Q-Network (DQN), proposto por Mnih et al. (2015), é o representante paradigmático dessa família, tendo demonstrado que redes neurais profundas são capazes de aproximar a Q-função em ambientes de alta dimensionalidade. Embora poderosos, métodos *value-based* operam intrinsecamente sobre espaços de ação discretos, a Q-função avalia cada ação possível e seleciona o máximo, tornando-os inadequados para o problema de alocação contínua de ativos,

no qual a ação é um vetor de pesos em $\mathbb{R}^n$ que não admite enumeração (MNIH et al., 2015).

Os métodos *Policy-Based*, por sua vez, parametrizam diretamente a política $\pi_\theta\left(\frac{a}{s}\right)$, geralmente uma rede neural com parâmetros θ , e a otimizam por gradiente, sem a intermediação de uma função de valor. Essa abordagem é naturalmente compatível com espaços de ação contínuos e de alta dimensionalidade, atendendo ao requisito do problema de alocação de portfólio. No entanto, a atualização dos parâmetros é guiada pelo retorno total do episódio, que apresenta alta variância em ambientes estocásticos, resultando em instabilidade e convergência lenta durante o treinamento, limitação que motivou diretamente o desenvolvimento de métodos mais robustos de gradiente de política (SUTTON; BARTO, 2018; SCHULMAN et al., 2015).

A arquitetura Actor-Critic emerge como síntese das duas famílias precedentes, combinando a parametrização direta da política com a estimação de uma função de valor que atua como mecanismo de redução de variância. Sua formulação, cujas origens remontam aos trabalhos de Barto, Sutton e Anderson (1983) e cuja análise de convergência foi estabelecida por Konda e Tsitsiklis (2000), organiza-se em dois componentes funcionalmente distintos que compõem uma única entidade decisora. O ator (actor) corresponde à política parametrizada $\pi_\theta\left(\frac{a}{s}\right)$, responsável pela seleção das ações e, portanto, pela interação efetiva com o ambiente. A crítica (critic) corresponde a uma função de valor de estado $V_\phi(s)$, que não emite ações, mas avalia a qualidade dos estados visitados, fornecendo ao ator um sinal de treinamento de menor variância que o retorno bruto do episódio (SUTTON; BARTO, 2018).

O mecanismo que articula os dois componentes é a função de vantagem, definida como $A(s_t, a_t) = Q(s_t, a_t) - V(s_t)$, cuja interpretação é direta: ela mensura o quanto a ação efetivamente executada superou, ou ficou aquém, do desempenho médio esperado para aquele estado. Ao substituir o retorno absoluto pela vantagem estimada no cálculo do gradiente da política, a arquitetura preserva o direcionamento correto da atualização e reduz substancialmente sua variância, mitigando a instabilidade que caracteriza os métodos policy-based puros (SCHULMAN et al., 2016). Cumpre observar que a crítica constitui maquinaria de treinamento, e não um segundo decisor: encerrada a fase de aprendizado, apenas a política do ator é retida para a inferência.

É nessa família que se inscrevem os algoritmos de controle contínuo do estado da arte, entre os quais o Advantage Actor-Critic (MNIH et al., 2016) e o Proximal Policy Optimization (SCHULMAN et al., 2017), adotado neste trabalho e detalhado na subseção seguinte. A arquitetura Actor-Critic reúne, assim, as duas propriedades exigidas pelo problema de alocação dinâmica de portfólio: a compatibilidade nativa com espaços de ação contínuos, herdada da abordagem policy-based, e a estabilidade de treinamento conferida pela estimação de valor, herdada da abordagem value-based.

2.3.3 O Algoritmo Proximal Policy Optimization (PPO)

O Proximal Policy Optimization (PPO) foi proposto por Schulman et al. (2017) como uma solução de primeira ordem para o principal problema dos métodos de gradiente de política, a instabilidade causada por atualizações excessivamente grandes dos parâmetros. Seu predecessor, o Trust Region Policy Optimization (TRPO), garantia atualizações seguras ao impor uma restrição de divergência KL entre a política corrente e a atualizada, mas o custo computacional das operações de segunda ordem, em particular, o cálculo da matriz Hessiana, inviabiliza sua aplicação em larga escala (SCHULMAN et al., 2015). O PPO atingiu desempenho equivalente ou superior ao TRPO com implementação dramaticamente mais simples, tornando-se o algoritmo de referência para problemas de controle contínuo em RL (SCHULMAN et al., 2017).

A inovação central do PPO é a função objetivo com *clipping*, que restringe a magnitude da atualização da política sem a necessidade de resolver um problema de

otimização com restrições. Formalmente, seja $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ a razão da probabilidade entre a nova política e a política que coletou os dados. O objetivo *clipped* é definido como:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right]$$

onde $\hat{A}_t$ é a estimativa da vantagem generalizada (Generalized Advantage Estimation - GAE) no instante t , e ϵ é um hiperparâmetro de *clipping* (tipicamente 0,1 ou 0,2). A operação *clip* impede que a razão $r_t(\theta)$ se afaste excessivamente de 1, delimitando o

tamanho efetivo do passo de atualização. O operador *min* assegura que o objetivo se torne um limite inferior (pessimista) do objetivo não *clipped*: quando a vantagem é positiva, a razão é limitada em $1+\epsilon$, impedindo que a política explore demais uma ação favorável; quando a vantagem é negativa, a razão é limitada em $1-\epsilon$, evitando que a política penalize excessivamente uma ação desfavorável com base em uma única amostra (SCHULMAN et al., 2017).

Do ponto de vista prático, o PPO realiza múltiplas épocas de atualização por minibatch sobre o mesmo conjunto de trajetórias coletadas, o que aumenta a eficiência amostral sem comprometer a estabilidade, vantagem que não era viável nos métodos de gradiente de política de primeira geração (SCHULMAN et al., 2017). A adequação do PPO ao problema de alocação contínua de ativos é corroborada pela literatura recente de finanças quantitativas: Uppaladinni et al. (2024) demonstram que a integração de RL com redes neurais para rebalanceamento em tempo real supera métodos determinísticos clássicos justamente pela capacidade do algoritmo de realizar múltiplas atualizações de política sobre os mesmos dados de mercado, adaptando-se a mudanças rápidas no regime econômico sem instabilidade de treinamento. A biblioteca Stable-Baselines3, utilizada neste trabalho, implementa o PPO com a política *MultiInputPolicy*, capaz de processar observações compostas por múltiplos espaços, neste caso, a janela de dados de mercado e o vetor de pesos correntes da carteira, e mapear esse estado enriquecido diretamente sobre logits contínuos no espaço de ação.

2.4 REBALANCEAMENTO DE PORTFÓLIO

A eficiência de um portfólio não está atrelada apenas à alocação inicial dos ativos, mas também, de acordo com Markowitz (1952), à manutenção da proporção entre ativos de risco e ativos livres de risco. Como cada ativo se valoriza e se desvaloriza em ritmos distintos, a estrutura inicial da carteira tende a se alterar ao longo do tempo. Nesse contexto, o rebalanceamento de carteira consiste na revisão periódica dos pesos dos ativos com o objetivo de retornar à alocação originalmente definida.

Ang (2014) explica que, sem o rebalanceamento, ativos mais arriscados e voláteis, mas com retornos históricos elevados, tendem a dominar a carteira, expondo

o investidor a riscos sistêmicos não planejados. Esse fenômeno é denominado deriva do portfólio.

As três principais estratégias de rebalanceamento definidas pela Vanguard Research (2019) são o rebalanceamento por calendário, o rebalanceamento por faixas de tolerância e o rebalanceamento por fluxo de caixa. Cada abordagem varia de acordo com os custos de transação e a sensibilidade às oscilações do mercado. Na primeira estratégia, os ajustes são realizados em intervalos fixos de tempo, independentemente das variações do mercado. Bernstein (2000) ressalta que esse método pode ignorar movimentos bruscos, embora apresente a vantagem de ser simples e reduzir decisões emocionais. Já o rebalanceamento por fluxo de caixa consiste em direcionar novos aportes para ativos abaixo da alocação-alvo ou realizar resgates prioritariamente nos ativos acima do peso desejado. Segundo Ang (2014), essa estratégia é a mais eficiente sob a ótica fiscal e de custos.

A estratégia utilizada pelo modelo treinado é o rebalanceamento por faixas de tolerância (thresholds), método que se baseia na volatilidade relativa dos ativos. De acordo com Swensen (2005), essa abordagem exige a definição de corredores em torno do alvo estratégico, de modo que o ajuste da carteira só é realizado quando o peso de um ativo ultrapassa o limite superior ou inferior da faixa estabelecida.

A escolha das faixas de tolerância é uma decisão importante nesse método, pois faixas muito estreitas geram maiores custos de transação, embora mantenham o portfólio mais balanceado, enquanto faixas muito amplas podem afastar a carteira do nível de risco desejado, conforme Ang (2014). Por essa razão, a aplicação de uma faixa de 5% foi considerada o equilíbrio ideal pela Vanguard Research (2019).

Entretanto, existem critérios específicos para a definição dessas faixas. De acordo com Maginn (2007), ativos mais voláteis necessitam de faixas mais amplas para evitar rebalanceamentos excessivos, enquanto ativos altamente correlacionados reduzem a necessidade de ajustes frequentes.

2.4.1 Custos do rebalanceamento

A execução de estratégias de rebalanceamento impõe custos operacionais e fiscais que podem comprometer os benefícios gerados pela diversificação e, conforme Maginn (2007), até consumir integralmente o benefício marginal do ajuste de pesos.

De acordo com Ang (2014), os custos associados ao rebalanceamento dividem-se em duas categorias principais: custos diretos (ou explícitos) e custos indiretos (ou implícitos).

Os custos explícitos são aqueles facilmente quantificáveis no momento da transação, como taxas de corretagem, emolumentos e tributação. Em mercados emergentes, como o mercado financeiro brasileiro, os custos de corretagem e as taxas de custódia e liquidação da bolsa de valores (B3) exercem um peso não negligenciável sobre carteiras. Além disso, de acordo com Swensen (2005), vender ativos que se valorizaram muito apenas para realizar o rebalanceamento da carteira pode gerar incidência de imposto sobre ganho de capital.

Os custos implícitos manifestam-se de forma mais sutil, o *bid-ask spread* representa a diferença entre o preço de oferta de compra e o de venda, em momentos de estresse ou em ativos de menor liquidez, esse diferencial se alarga, tornando o custo maior (HASBROUCK, 2007). O impacto de mercado, por sua vez, refere-se ao deslocamento de preço causado pela própria ordem do investidor, ordens de venda de grande volume pressionam o preço para baixo, enquanto grandes ordens de compra elevam o preço (ALMGREN; CHRISS, 2000).

2.5 CONCEITOS ECONÔMICOS E MÉTRICAS FINANCEIRAS

Para a correta modelagem do ambiente de treinamento e da interpretação dos resultados, é necessário definir os conceitos financeiros e as variáveis macroeconômicas que compõem esse estudo.

2.5.1 Exchange Traded Funds (ETFs)

Exchange Traded Funds (ETFs) são fundos negociados em bolsa que replicam índices de referência (benchmark), permitindo diversificação instantânea, liquidez e alocação macroestratégica (GASTINEAU, 2010).

2.5.2 Variáveis Macroeconômicas

A taxa SELIC é a taxa de juros média praticada nas operações compromissadas com títulos públicos federais com prazo de um dia útil. É o principal

instrumento de política monetária do Banco Central do Brasil (BC), cuja meta é definida pelo Comitê de Política Monetária (Copom) (BCB, [1.1]). Em sua função de taxa livre de risco da economia brasileira, a SELIC serve como denominador do prêmio de risco exigido por investidores para assumir exposição a ativos mais voláteis (BCB, 2024).

O IPCA (Índice Nacional de Preços ao Consumidor Amplo) é o indicador oficial de inflação do país, calculado pelo IBGE (Instituto Brasileiro de Geografia e Estatística). Tem por objetivo medir a variação de preços de um conjunto de produtos e serviços consumidos no varejo por famílias com rendimento mensal entre 1 e 40 salários-mínimos (IBGE, 2024).

Variáveis macroeconômicas como taxa de juros e inflação exercem influência sistemática e documentada sobre os preços de diferentes classes de ativos, justificando sua inclusão no espaço de estados do modelo (BARTKUS; PALEVICIENE, 2013)

2.5.3 Métricas de Avaliação de Desempenho

A avaliação quantitativa de estratégias de investimento requer métricas que capturem tanto o retorno gerado quanto o risco assumido para gerá-lo. A análise restrita ao retorno absoluto é insuficiente, pois ignora a trajetória e a severidade das perdas (SEN, 2022).

A volatilidade, mensurada pelo desvio-padrão dos retornos, constitui a medida mais elementar de risco de um ativo ou carteira, quantificando a dispersão dos retornos em torno de sua média. Por mensurar a dispersão total dos retornos, a volatilidade trata simetricamente ganhos e perdas, limitação que motiva o recurso a métricas complementares, como o Índice de Sortino (SEN, 2022).

O Índice de Sharpe quantifica o retorno excedente por unidade de risco total (volatilidade), sendo calculado como:

$$S = \frac{R_p - R_f}{\sigma_p}$$

onde R_p é o retorno médio do portfólio, R_f é a taxa livre de risco (neste trabalho, a SELIC diária anualizada) e σ_p é o desvio padrão anualizado dos retornos.

O índice penaliza a volatilidade total sendo mais adequado para distribuições simétricas de retornos (SEN, 2022). Ou seja, o índice indica quanto retorno adicional é obtido para cada unidade de risco assumida, sendo muito utilizado para comparar a relação risco-retorno do ativo.

O Índice de Sortino aprimora o Índice de Sharpe ao calcular o retorno excedente por unidade de risco de queda (*downside deviation*), penalizando apenas a volatilidade negativa:

$$Sortino = \frac{R_p - R_f}{\sigma_d}$$

Em distribuições assimétricas, comuns em ativos de renda variável, o Sortino é uma medida mais informativa do risco efetivamente suportado pelo investidor (SEN, 2022), oferecendo um dado mais sensível e apropriado quando o objetivo é penalizar apenas a volatilidade desfavorável.

O Maximum Drawdown (MDD) representa a maior queda percentual acumulada observada entre um pico e um fundo subsequente durante o período analisado:

$$MDD = \min_t \left[\frac{V_t - \max_{\tau \leq t} V_\tau}{\max_{\tau \leq t} V_\tau} \right]$$

onde V_t é o valor do portfólio no instante t .

O MDD é uma medida de risco de queda (*tail risk*) que mede a resiliência do ativo. Essa métrica tem especial relevância pois captura o pior cenário experimentado ao longo da trajetória do portfólio, independentemente do retorno final (SEN, 2022).

O Índice de Calmar relaciona o retorno anualizado com o Maximum Drawdown:

$$Calmar = \frac{R_a}{|MDD|}$$

onde R_a representa o retorno anualizado do portfólio.

Por penalizar diretamente as perdas máximas, o Calmar é particularmente adequado à avaliação de estratégias que operam em ambientes de alta volatilidade, como o mercado financeiro brasileiro durante períodos de estresse macroeconômico (SEN, 2022).

2.6 FUNÇÃO DE RECOMPENSA

A função de recompensa constitui o sinal escalar que o modelo busca maximizar ao longo do treinamento, traduzindo o objetivo do problema em um critério de otimização e orientando toda a política aprendida. Sua especificação é uma das decisões de desenho mais determinantes em aplicações de Aprendizado por Reforço a finanças, e a literatura registra um espectro de formulações. As abordagens baseadas no retorno distinguem-se, em primeiro lugar, pelo tratamento que conferem à fricção transacional. Modelos orientados pela maximização do retorno bruto, que ignoram os custos de operação, tendem a desenvolver políticas hiperativas e financeiramente insustentáveis (FERREIRA, 2025). Uma estratégia consolidada para contornar essa falha consiste em incorporar o custo de transação diretamente ao sinal de recompensa, de modo que o modelo internalize o ônus financeiro imediato de cada operação (JIANG; XU; LIANG, 2017).

No que tange ao risco, por sua vez, parte da literatura propõe recompensas que penalizam explicitamente as perdas de pico a vale: Almahdi e Yang (2017) demonstram que uma função baseada no Maximum Drawdown superou, na gestão de portfólios de ETFs, uma recompensa baseada no índice de Sharpe, sugerindo que tal penalização favorece a robustez da política em cenários adversos. O presente trabalho situa-se entre essas vertentes: adota uma recompensa fundamentada no log-retorno líquido de custos, que responde diretamente ao problema da hiperatividade, ao fazer o modelo sentir a fricção transacional, porém de caráter deliberadamente minimalista, delegando a contenção do risco de queda à modelagem do ambiente, em vez de penalizá-lo no próprio sinal. A justificativa dessa escolha e o reconhecimento de suas implicações sobre a robustez são desenvolvidos na seção 3.4.3 e retomados entre as direções de trabalho futuro.

2.7 TRABALHOS RELACIONADOS

A aplicação de técnicas de Aprendizado por Reforço ao problema de alocação de portfólio tem ganhado crescente tração na literatura de finanças quantitativas, especialmente após a consolidação das arquiteturas de Deep RL. A revisão a seguir organiza-se em torno dos eixos que informaram diretamente as decisões de projeto

desta pesquisa: a arquitetura da política, a escolha algorítmica, o desenho da função de recompensa, a modelagem da fricção transacional e a incorporação de contexto macroeconômico ao espaço de estados.

A contribuição seminal para a gestão contínua de portfólio deve-se a Jiang, Xu e Liang (2017), que propuseram o framework EIIE (Ensemble of Identical Independent Evaluators), uma das primeiras arquiteturas de Deep RL concebidas especificamente para o problema, demonstrando desempenho superior a estratégias convencionais de alocação em dados de criptoativos. A contribuição evidenciou a viabilidade da abordagem, porém a ausência de modelagem explícita de custos de transação e de variáveis macroeconômicas no espaço de estados constitui limitação estrutural relevante, sobretudo quando se pretende transpor o método a mercados regulamentados, nos quais a fricção transacional consome fração proporcionalmente maior do retorno bruto.

Almahdi e Yang (2017) avançaram nessa direção ao desenvolver um sistema adaptativo de negociação baseado em RL recorrente com função de recompensa orientada ao Maximum Drawdown esperado, demonstrando que o foco no risco de queda produz políticas mais robustas do que a maximização direta do índice de Sharpe em carteiras de ETFs. A abordagem influenciou diretamente o desenho do presente trabalho, em particular a eleição de métricas de risco de queda, os índices de Sortino e de Calmar, como indicadores primários de qualidade da política aprendida.

A transposição da abordagem para mercados acionários regulamentados foi investigada por Liang et al. (2018), que implementaram comparativamente três algoritmos de controle contínuo, Deep Deterministic Policy Gradient, Proximal Policy Optimization e Policy Gradient, aplicados à gestão de portfólio no mercado acionário chinês. O achado dos autores é contraintuitivo e metodologicamente determinante para esta pesquisa: o método de gradiente de política simples revelou-se mais adequado ao ambiente financeiro do que o DDPG e o PPO, a despeito da sofisticação teórica superior destes últimos. O resultado sugere que as garantias de estabilidade que motivam a adoção do PPO em domínios de controle robótico, nos quais a dinâmica do ambiente é regida por leis determinísticas e a razão sinal-ruído é elevada, não se transferem automaticamente ao domínio financeiro, caracterizado por

estocasticidade acentuada e baixa previsibilidade estrutural. A evidência fundamenta a postura cautelosa adotada neste trabalho, que não pressupõe a superioridade do algoritmo escolhido, mas a submete a verificação empírica sob regimes macroeconômicos distintos.

A relação entre variáveis macroeconômicas e o mercado acionário brasileiro, por sua vez, foi investigada por Theodoro (2018) mediante modelos de Vetores Autorregressivos, tendo o autor concluído que a associação entre SELIC, IPCA e o IBOVESPA é estatisticamente inconclusiva sob uma ótica estritamente linear. O achado é determinante para o posicionamento desta pesquisa. Ele não indica ausência de relação, mas sugere que a interação entre a macroeconomia e a precificação de ativos possui natureza intrinsecamente não-linear, insuscetível de captura por especificações paramétricas lineares. Justifica-se, assim, o recurso a um aproximador de função não-linear, a rede neural que parametriza a política, e a incorporação direta dos indicadores macroeconômicos ao espaço de estados, em vez de sua modelagem exógena como variáveis explicativas de um sistema de equações.

No contexto brasileiro de aplicação de Deep RL, Ferreira (2025) aplicou o algoritmo PPO a uma carteira de criptoativos, documentando que modelos treinados sem variáveis macroeconômicas no estado e sem modelagem de custos de transação desenvolveram políticas hiperativas e financeiramente insustentáveis. O presente trabalho posiciona-se como extensão metodológica direta dessa investigação, ao transferi-la ao mercado de ETFs regulamentados, incorporar o contexto macroeconômico ao espaço de estados e modelar explicitamente a fricção transacional no ambiente de simulação.

A revisão conduzida por Gunjan e Bhattacharyya (2023) sobre técnicas de otimização de portfólio com inteligência artificial corrobora a pertinência dessas escolhas, identificando a ausência de modelagem de custos e de contexto macroeconômico como as lacunas mais frequentes nos estudos analisados. A Tabela 2 sintetiza o posicionamento comparativo dos trabalhos revisados segundo os eixos que estruturam esta pesquisa.

Tabela 2 – Posicionamento comparativo dos trabalhos relacionados

Trabalho	Mercado	Abordagem	Custos de transação	Contexto macroeconômico	Avaliação multi-regime
Jiang, Xu e Liang (2017)	Criptoativos	Deep RL (EIIE)	Não	Não	Não
Almahdi e Yang (2017)	ETFs (EUA)	RL recorrente	Sim	Não	Não
Liang et al. (2018)	Ações (China)	Deep RL (DDPG, PPO, PG)	Sim	Não	Não
Theodoro (2018)	Ações (Brasil)	VAR (linear)	Não se aplica	Sim (exógeno)	Não
Ferreira (2025)	Criptoativos	Deep RL (PPO)	Não	Não	Não
Presente trabalho	ETFs (Brasil)	Deep RL (PPO)	Sim	Sim (no estado)	Sim (três regimes)

Fonte: Elaboração própria (2026)

A leitura do quadro permite delimitar a contribuição desta pesquisa em três frentes. A primeira, de natureza empírica, consiste na aplicação do arcabouço a uma carteira multiclasse de ETFs do mercado brasileiro, ambiente cuja estrutura de juros elevados e cuja sensibilidade a choques macroeconômicos o diferenciam substancialmente dos mercados norte-americano e chinês que predominam na literatura revisada. A segunda reside na integração simultânea da fricção transacional e do contexto macroeconômico ao ambiente de simulação, combinação que, conforme evidencia o quadro, nenhum dos trabalhos precedentes contempla, e cuja ausência Gunjan e Bhattacharyya (2023) identificam como a lacuna mais recorrente do campo. A terceira, e possivelmente a mais consequente, é de natureza metodológica: ao submeter a política aprendida a um protocolo multi-seed sob três regimes macroeconômicos distintos, em vez de reportar o desempenho de uma execução única em um período único, esta pesquisa desloca a questão de investigação da constatação de superioridade para a delimitação das condições sob as quais a superioridade se verifica. O achado de robustez condicional ao regime, detalhado no Capítulo 4, dialoga diretamente com a cautela algorítmica sinalizada por Liang et al. (2018) e é indissociável desse desenho experimental, não sendo observável sob o protocolo de avaliação convencional adotado nos trabalhos precedentes.

3 METODOLOGIA

Este capítulo descreve o método adotado para o desenvolvimento e a avaliação do modelo de Aprendizado por Reforço proposto neste trabalho. A exposição parte da constituição da base de dados e da definição do universo de investimento, avança pela formalização do ambiente de simulação e pela configuração do modelo, e culmina no protocolo de avaliação empregado para confrontar a política aprendida com estratégias de referência.

3.1 VISÃO GERAL DO FLUXO METODOLÓGICO

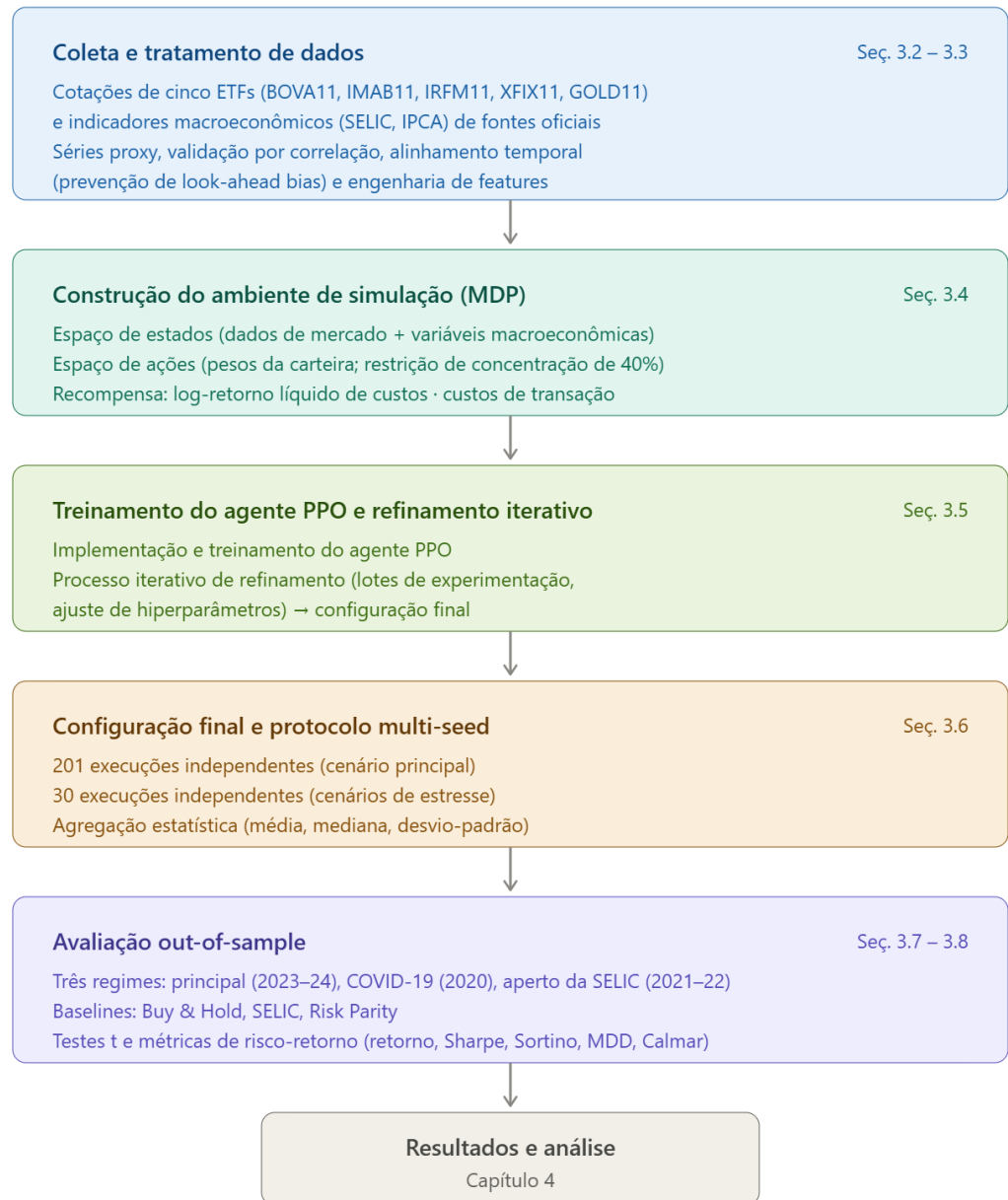
O desenvolvimento do trabalho organiza-se em quatro etapas sequenciais. A primeira compreende a coleta e o tratamento dos dados, na qual se constitui uma base histórica íntegra e temporalmente alinhada, integrando cotações de fundos de índice a indicadores macroeconômicos. A segunda consiste na construção do ambiente de simulação, responsável por formalizar o problema de alocação como um Processo de Decisão de Markov e por reproduzir ao máximo as dinâmicas do mercado financeiro brasileiro. A terceira etapa abrange o treinamento do modelo, conduzido de maneira iterativa e com melhorias a cada ciclo. A quarta e última etapa corresponde à avaliação, na qual o desempenho da política aprendida é confrontado com estratégias de referência em distintos regimes de mercado.

Cabe destacar que o desenvolvimento do modelo não se deu por meio de uma única execução de treinamento, mas mediante um processo experimental incremental, organizado em lotes sucessivos. Cada lote introduziu uma modificação pontual e controlada, seja na configuração de hiperparâmetros, na função de recompensa ou na dinâmica de transação do ambiente, cujo efeito foi mensurado isoladamente antes de orientar a configuração subsequente. Essa abordagem, detalhada na seção 3.5, aproxima a construção do modelo do método experimental, na medida em que cada decisão de projeto é justificada empiricamente pelos resultados observados na iteração anterior, e não por escolhas arbitrárias a priori.

A Figura 1 apresenta uma visão geral do fluxo metodológico adotado nesta pesquisa, desde a coleta e o tratamento dos dados até a avaliação do

modelo em diferentes regimes de mercado, etapas detalhadas nas seções subsequentes deste capítulo.

Figura 1 – Fluxo Metodológico da Pesquisa



Fonte: Elaboração própria (2026)

3.1.1 Ambiente Computacional e Stack Tecnológica

A totalidade dos experimentos foi conduzida em ambiente local, sem recurso a infraestrutura de nuvem. A máquina utilizada é um notebook equipado com processador Intel Core i7-10750H (6 núcleos físicos, frequência base de 2,60 GHz), 16 GB de memória RAM e placa gráfica dedicada NVIDIA GeForce GTX 1660 Ti com 6 GB de memória, operando sob o sistema Windows 11 de 64 bits. Os modelos treinados foram armazenados localmente em unidade de estado sólido (SSD), e o versionamento do código-fonte foi gerido por meio da ferramenta Git, assegurando o rastreamento das alterações ao longo do processo iterativo de desenvolvimento descrito na seção 3.5.

A implementação foi desenvolvida na linguagem Python. O treinamento do modelo apoiou-se na biblioteca Stable-Baselines3¹ (versão 2.2.1 ou superior), que fornece uma implementação consolidada e validada do algoritmo Proximal Policy Optimization e que usa a biblioteca PyTorch para treinamento dos modelos, instalada como dependência. A interface entre o modelo e o simulador de mercado seguiu o padrão estabelecido pela biblioteca Gymnasium² (versão 0.29.1 ou superior), responsável pela especificação dos espaços de observação e de ação do ambiente customizado. O manuseio e a transformação das séries temporais foram realizados com as bibliotecas Pandas³ (versão 2.2.0 ou superior) e NumPy⁴ (versão 1.26.3 ou superior), enquanto a geração das visualizações empregou Matplotlib⁵ e Seaborn⁶. A coleta de dados recorreu às bibliotecas yfinance⁷, para as cotações de mercado, e python-bcb⁸, para o acesso aos dados de variáveis macroeconômicas.

¹ [Documentação Stable-Baselines3](#)

² [Documentação Gymnasium](#)

³ [Documentação Pandas](#)

⁴ [Documentação Numpy](#)

⁵ [Documentação Matplotlib](#)

⁶ [Documentação Seaborn](#)

⁷ [Documentação yFinance](#)

⁸ [Documentação python-bcb](#)

3.2 DEFINIÇÃO DO UNIVERSO DE ATIVOS

A primeira etapa do trabalho consistiu na definição do universo de investimento sobre o qual o modelo atuará. Optou-se pela utilização exclusiva de ETFs (*Exchange Traded Funds*), em detrimento de ações individuais, por três razões fundamentais. Em primeiro lugar, os ETFs proporcionam diversificação instantânea dentro de cada classe de ativos, reduzindo o risco inerente associado a um único emissor. Em segundo, sua estrutura permite que o modelo opere no nível da alocação macroestratégica entre classes (renda variável, renda fixa, fundos imobiliários e proteção) em vez de exigir a análise individual de empresas específicas. Por fim, a negociação em bolsa confere liquidez e transparência de preços compatíveis com as condições de execução pressupostas pelo ambiente de simulação.

A seleção dos ativos específicos pautou-se por três critérios. O critério de representatividade exigiu que cada ETF replicasse um índice de referência consolidado de uma classe distinta, de modo que o conjunto resultante oferecesse exposição abrangente às principais categorias do mercado financeiro brasileiro. O critério de liquidez priorizou os fundos de maior volume de negociação em suas respectivas categorias, garantindo que a simulação de compra e venda refletisse condições realistas de execução. O critério de profundidade histórica e disponibilidade de dados, por fim, exigiu que cada classe de ativo pudesse ser representada por uma série temporal de extensão suficiente para abranger múltiplos regimes de mercado, obtida a partir de fontes públicas e de acesso gratuito. Nos casos em que o ETF selecionado não dispunha de cotações para a totalidade do período de análise, recorreu-se a séries proxy de igual acessibilidade, cuja aderência foi validada conforme detalhado na seção 3.3.

A aplicação desses critérios resultou em uma cesta composta por cinco ETFs, cada um representando uma classe de ativos com função estratégica definida na carteira. O BOVA11 fornece exposição à renda variável, replicando o Ibovespa, principal índice do mercado acionário brasileiro (B3, 2024). O XFIX11 representa o segmento de fundos de investimento imobiliário, acompanhando o IFIX, índice que reúne os fundos imobiliários mais negociados na bolsa brasileira. Os ativos IMAB11 e IRFM11 compõem a parcela de renda fixa, replicando, respectivamente, o índice IMA-B, carteira de títulos públicos federais indexados à inflação (Tesouro IPCA+), e o

índice IRF-M, carteira de títulos públicos prefixados (Tesouro Prefixado), ambos calculados pela ANBIMA (ANBIMA, 2024). Por fim, o GOLD11 atua como o ativo de proteção do portfólio. Embora a literatura consolide o ouro como um porto seguro clássico durante crises financeiras em países desenvolvidos (BAUR; McDERMOTT, 2010), no mercado brasileiro, embora um mercado ainda emergente, a proteção se dá pela estrutura operacional do fundo. Como o ETF não possui proteção cambial, sua cotação embute a variação do dólar americano, ativo que historicamente atua como um robusto *hedge* contra quedas no mercado acionário local em momentos de fuga de capital (SENDACZ, 2015). Dessa forma, o GOLD11 confere à carteira uma dupla camada de proteção: a resiliência do metal somada à natural desvalorização do real.

O conjunto de dados resultante abrange o período de fevereiro de 2015 a dezembro de 2024, totalizando 2.575 pregões após o tratamento descrito na seção seguinte. A extensão do histórico, aproximadamente dez anos, é particularmente relevante por contemplar regimes de mercado heterogêneos, incluindo períodos de estabilidade, a crise associada à pandemia de COVID-19 e o ciclo de elevação acentuada da taxa básica de juros, condições exploradas na avaliação de robustez descrita na seção 3.7.

3.3 COLETA E TRATAMENTO DE DADOS

A constituição da base de dados constitui etapa crítica em trabalhos de aprendizado de máquina aplicados a finanças, uma vez que a qualidade e a integridade temporal das séries implicam diretamente na validade das políticas aprendidas pelo modelo. O pipeline de dados foi implementado de forma programática e reproduzível, abrangendo a coleta das séries em fontes públicas, a construção de séries representativas para os ativos sem histórico completo, o alinhamento temporal entre dados de mercado e indicadores macroeconômicos e a derivação das variáveis que compõem o espaço de estados do modelo.

O período de análise compreende o intervalo de janeiro de 2015 a dezembro de 2024, totalizando aproximadamente dez anos de cotações diárias. Essa janela foi deliberadamente escolhida por contemplar regimes macroeconômicos heterogêneos, incluindo ciclos monetários diferentes, como o choque associado à pandemia de

COVID-19, condição necessária para a avaliação da capacidade de generalização do modelo em diferentes contextos de mercado. As cotações dos fundos de índice foram obtidas junto ao serviço yfinance, ao passo que os indicadores macroeconômicos e os índices de renda fixa foram coletados diretamente do Sistema Gerenciador de Séries Temporais (SGS) do Banco Central do Brasil, por meio da biblioteca python-bcb.

A coleta foi estruturada de modo incremental: o pipeline verifica a existência de uma base previamente armazenada e, em caso positivo, requisita às fontes apenas as observações posteriores à última data registrada, concatenando-as ao histórico existente. Esse mecanismo evita requisições redundantes às APIs externas e viabiliza a atualização eficiente da base à medida que novos dados se tornam disponíveis. As séries brutas resultantes são persistidas localmente antes de qualquer transformação, preservando-se a rastreabilidade entre os dados originais e a base processada que efetivamente alimenta o ambiente de simulação.

3.3.1 Construção de séries Proxy e validação por Correlação

Um desafio na constituição de bases históricas para o mercado brasileiro reside na juventude relativa de muitos fundos de índice. Embora o período de análise se estenda desde 2015, nem todos os ETFs que compõem o universo de investimento dispunham de cotações ao longo de todo o intervalo, em virtude de suas datas de listagem mais recentes. Para preservar a extensão da base, condição crucial para a exposição do modelo a múltiplos regimes de mercado, recorreu-se à construção de séries proxy para os ativos com histórico incompleto. Nos ramos de pesquisa, uma variável proxy é uma variável empregada em lugar de uma variável de interesse que não pode ser diretamente medida ou cujos dados são indisponíveis no intervalo temporal de análise (WOOLDRIDGE, 2016). Para que uma proxy seja considerada adequada, exige-se que ela apresente forte correlação com a variável-alvo em um período de sobreposição entre as séries, capturando seu comportamento de modo a permitir inferência estatística válida sobre o fenômeno de interesse (CLINTON, 2011). Em todos os casos, a adequação do proxy foi submetida à validação estatística mediante o cálculo do coeficiente de correlação de Pearson entre a série substituta e

o ativo-alvo, no período de sobreposição em que ambos dispunham de dados (2021–2024).

Três mecanismos distintos de construção foram empregados, conforme a natureza da indisponibilidade de cada ativo. O primeiro mecanismo consiste na substituição integral por um ativo correlato. O ETF XFIX11, representante da classe de fundos imobiliários, não possui histórico que recue até 2015; em seu lugar, adotou-se o KNRI11, um fundo imobiliário de tijolo de perfil semelhante e histórico mais longo, como representante da classe ao longo de todo o período analisado. A pertinência dessa substituição é sustentada pela correlação de 0,9486 entre os retornos dos dois ativos no período de sobreposição, indicativa de que o KNRI11 reproduz com fidelidade o comportamento do segmento imobiliário que o XFIX11 busca capturar.

O segundo mecanismo consiste na construção de uma série sintética a partir de ativos primários. O ETF GOLD11, que oferece exposição ao ouro denominado em reais, não existia no início do período de análise. Para reproduzir seu comportamento, construiu-se uma série sintética pela multiplicação da cotação do ETF internacional GLD (SPDR Gold Shares, que replica o preço do ouro em dólares) pela taxa de câmbio entre o dólar e o real ($BRL=X$), obtendo-se assim uma representação do ouro em moeda local independentemente da existência do ativo no mercado brasileiro. A validação dessa construção, comparando a série sintética ao GOLD11 real no período de sobreposição, resultou em correlação de 0,9734.

O terceiro mecanismo consiste na costura temporal entre uma fonte primária histórica e o ETF correspondente. Os ativos de renda fixa IMAB11 e IRFM11 replicam, respectivamente, os índices IMA-B e IRF-M, calculados e disponibilizados oficialmente pelo Banco Central do Brasil por meio do SGS (séries 12466 e 12462). Esses índices oficiais cobrem o período histórico anterior à listagem dos ETFs correspondentes. Como, contudo, as séries oficiais deixam de ser atualizadas a partir de determinado ponto, aplicou-se um procedimento de costura (implementado na rotina `stitch_series`): identifica-se a última observação com variação não nula na série do Banco Central e, a partir desse ponto de junção, a série do ETF obtida via `yfinance` é reescalada para preservar a continuidade de nível de preços, sendo então concatenada ao histórico oficial. A validação no período de sobreposição confirmou a aderência de ambas as séries, com correlação de 0,9862 para o par IMAB11 / IMA-B e de 0,9033 para o par IRFM11 / IRF-M.

A Tabela 3 sintetiza os procedimentos de construção de proxy e os respectivos coeficientes de validação. A magnitude das correlações obtidas, todas superiores a 0,90, sustenta a adequação das séries substitutas como representantes fiéis dos ativos-alvo, conferindo robustez à base histórica empregada no treinamento e na avaliação do modelo.

Tabela 3 – Construção de Séries Proxy e Validação por Correlação

Ativo Alvo	Mecanismo	Série utilizada	Correlação de Pearson
XFIX11	Substituição Integral	KNRI11	0,9486
GOLD11	Série sintética	GLD × BRL (USD/BRL)	0,9734
IMAB11	Costura da série	Índice IMA-B + ETF	0,9862
IRFM11	Costura da série	Índice IRF-M + ETF	0,9033

Fonte: Elaboração própria com base na análise empírica (2026)

3.3.2 Alinhamento Temporal e Prevenção de Look-Ahead Bias

A integração de séries provenientes de fontes heterogêneas, cotações de mercado, indicadores macroeconômicos e índices de renda fixa exige um cuidadoso alinhamento temporal, sob pena de introduzir inconsistências que comprometam a validade da simulação. Dois problemas distintos foram tratados nesta etapa: a harmonização dos calendários das diferentes séries e a prevenção do viés de antecipação (look-ahead bias).

O primeiro problema decorre da ausência de coincidência entre os calendários das fontes. Enquanto as cotações dos ETFs seguem o calendário de pregões da bolsa, os indicadores macroeconômicos obedecem a periodicidades próprias de divulgação. Para harmonizar as séries, o conjunto integrado foi reindexado sobre um calendário regular de dias úteis, e as observações eventualmente ausentes foram preenchidas pela técnica de propagação do último valor válido (forward fill). Essa abordagem assegura que, em cada dia de decisão, o modelo disponha de um valor definido para todas as variáveis de estado, refletindo a informação mais recente efetivamente disponível até aquele momento, sem recorrer a qualquer forma de interpolação que utilizasse dados futuros.

O segundo problema é o viés de antecipação. Esse viés ocorre quando o modelo tem acesso, em um dado instante de decisão, a uma informação que somente

se tornaria disponível em momento posterior, conferindo-lhe uma vantagem irreal e inflando artificialmente seu desempenho na avaliação. No presente trabalho, a fonte mais evidente desse risco reside no IPCA, cujo valor referente a um determinado mês é divulgado oficialmente pelo IBGE apenas semanas após o encerramento do período de referência. Utilizar o índice na data de referência, e não na data de sua efetiva divulgação, equivaleria a permitir que o modelo conhecesse a inflação antes de sua publicação. Para eliminar esse risco, a série do IPCA foi deslocada em 1 mês, ou seja, o modelo recebe os dados da inflação de janeiro apenas no mês de fevereiro, simulando a defasagem entre a competência do índice e sua disponibilização pública, de modo que o modelo apenas incorpore a informação inflacionária no momento em que ela seria realmente conhecida por um investidor.

Por fim, a base integrada e alinhada foi particionada cronologicamente em três subconjuntos disjuntos, respeitando a ordem temporal das observações para preservar a integridade da avaliação fora da amostra. O conjunto de treinamento compreende as observações anteriores a janeiro de 2022. O conjunto de validação, utilizado para a seleção do melhor modelo durante o treinamento, abrange o intervalo entre janeiro de 2022 e dezembro de 2022, e o conjunto de teste, reservado exclusivamente para a avaliação final out-of-sample, corresponde ao período a partir de janeiro de 2023. Essa segregação estritamente temporal, em oposição a particionamentos aleatórios, inadequados a séries temporais financeiras, garante que o desempenho reportado reflita a capacidade genuína de generalização do modelo sobre dados jamais vistos durante o aprendizado.

3.3.3 Engenharia de Features

A representação direta de preços ou retornos brutos não é, isoladamente, capaz de fornecer ao modelo um sinal informativo sobre o regime de mercado em curso. Para que o modelo disponha de uma caracterização estatística adequada do comportamento recente de cada ativo, foram derivadas quatro variáveis para cada um dos cinco ETFs, totalizando vinte indicadores específicos da carteira, acrescidos das duas variáveis macroeconômicas que compõem o estado completo. O cálculo dessas variáveis ocorre sobre uma janela móvel de vinte e um pregões, equivalente a aproximadamente um mês de negociação.

A primeira variável é o log-retorno diário, definido pela razão logarítmica entre o preço de fechamento corrente e o do pregão anterior:

$$r_{i,t} = \ln\left(\frac{P_{it}}{P_{i,t-1}}\right)$$

A adoção do retorno logarítmico (log-retorno), em detrimento do retorno percentual simples, fundamenta-se em uma propriedade matemática essencial para o treinamento do modelo: a aditividade no tempo. Diferentemente dos retornos simples, cuja acumulação exige o cálculo de produtos contínuos, a soma de sucessivos log-retornos ao longo de um horizonte t equivale exatamente à variação real entre o capital final e o inicial, expressa formalmente por $\ln(V_t - V_0)$ (CAMPBELL; LO; MACKINLAY, 1997). Na prática, isso garante que, ao configurar a função de recompensa para maximizar a soma contínua dos log-retornos (conforme detalhado na seção 3.4.3), o modelo de Aprendizado por Reforço otimiza diretamente a taxa de crescimento composto do portfólio. Adicionalmente, essa modelagem protege o modelo contra distorções estatísticas, visto que a média aritmética de retornos simples tende a superestimar o desempenho efetivo da carteira em cenários de alta volatilidade (HUDSON; GREGORIOU, 2015).

A segunda variável é a volatilidade móvel, calculada como o desvio padrão amostral dos log-retornos sobre a janela de vinte e um dias:

$$\sigma_{i,t} = \sqrt{\frac{1}{W-1} \sum_{k=t-W+1}^t (r_{i,k} - \bar{r}_{i,t})^2}$$

A volatilidade móvel oferece ao modelo uma medida do risco recente associado a cada ativo, viabilizando a discriminação entre períodos de mercado calmos e turbulentos.

A terceira variável é o índice de Sharpe móvel, que sintetiza a relação entre o retorno excedente médio e o risco total assumido no período de cálculo:

$$Sharpe_{i,t} = \frac{\bar{r}_{i,t} - \bar{r}_{f,r}}{\sigma_{i,t}}$$

onde $\bar{r}_{i,t}$ é a média móvel dos log-retornos do ativo i sobre a janela de vinte e um dias, $\bar{r}_{f,t}$ é a média móvel da SELIC diária no mesmo período, utilizada como taxa livre de risco, e $\sigma_{i,t}$ é a volatilidade móvel definida anteriormente. Diferentemente do uso

clássico do índice de Sharpe sobre o retorno total de uma série, a forma móvel adotada neste trabalho fornece ao modelo uma estimativa local da eficiência risco-retorno de cada ativo, atualizada a cada pregão. Para conferir robustez numérica ao indicador em situações de volatilidade próxima de zero, o denominador é estabilizado por uma constante $\varepsilon = 10^{-8}$, e o valor resultante é saturado no intervalo $[-5, +5]$, evitando que valores extremos, gerados em períodos de baixíssima volatilidade, propaguem instabilidade ao processo de aprendizado.

A quarta variável é o índice de Sortino móvel, que aprimora o índice de Sharpe ao substituir a volatilidade total pela volatilidade de queda (downside deviation), penalizando apenas o risco assimétrico decorrente de retornos negativos:

$$\text{Sortino}_{i,t} = \frac{\bar{r}_{i,t} - \bar{r}_{f,t}}{\sigma_{d,i,t}}$$

onde a volatilidade de queda $\sigma_{d,i,t}$ é calculada como a raiz quadrada da média dos quadrados dos log-retornos negativos sobre a janela móvel, com os retornos positivos truncados em zero antes da agregação. Assim como no índice de Sharpe, o resultado é saturado no intervalo $[-5, +5]$ para fins de estabilidade numérica. A inclusão simultânea do Sharpe e do Sortino no espaço de estados é proposital, fornecendo ao modelo duas perspectivas complementares sobre a qualidade do desempenho recente de cada ativo. Enquanto o primeiro penaliza a volatilidade total, incluindo flutuações positivas, o segundo penaliza exclusivamente o risco de perdas.

Por fim, as duas variáveis macroeconômicas, a taxa SELIC diária e o IPCA com a defasagem temporal descrita na seção 3.3.2, compõem o estado sem qualquer transformação adicional, sendo simplesmente alinhadas ao calendário de negociação e propagadas conforme já descrito. Essas variáveis fornecem ao modelo a referência do regime macroeconômico vigente, em particular, o nível da taxa básica de juros, que estabelece o custo de oportunidade da renda variável, e a inflação corrente, que afeta diretamente a rentabilidade real dos ativos indexados.

3.4 CONSTRUÇÃO DO AMBIENTE DE SIMULAÇÃO

O núcleo metodológico deste trabalho reside na construção de um ambiente de simulação que formaliza o problema de alocação de portfólio como um Processo de Decisão de Markov, permitindo que o modelo de Aprendizado por Reforço interaja

com uma representação realista do mercado financeiro. O ambiente foi implementado em Python, utilizando a interface padronizada da biblioteca Gymnasium, que estabelece o protocolo de comunicação entre o modelo e o ambiente por meio dos métodos de reinicialização (*reset*) e de transição (*step*). A formalização que se segue detalha os quatro elementos que constituem o Processo de Decisão de Markov subjacente: o espaço de estados, o espaço de ações, a função de recompensa e a dinâmica de transição com seus respectivos custos.

A escolha de desenvolver um ambiente customizado, em detrimento da adoção de simuladores pré-existentes, justifica-se pela necessidade de incorporar elementos específicos do problema de pesquisa, notadamente a integração de variáveis macroeconômicas ao espaço de estados e a modelagem explícita da fricção transacional, que não são contemplados por simuladores genéricos. O capital inicial de cada episódio foi fixado em R\$ 10.000,00, e a carteira é inicializada com pesos uniformemente distribuídos entre os cinco ativos, refletindo uma alocação neutra na ausência de informação prévia.

3.4.1 Espaço de Estados

O espaço de estados foi concebido para fornecer ao modelo uma representação rica e contextualizada das condições de mercado em cada instante de decisão. Em vez de expor o modelo a um único vetor instantâneo de preços, optou-se por uma representação estruturada que combina uma janela temporal de dados de mercado com a posição corrente da carteira, materializada na forma de um espaço de observação composto (*Dict*), com dois componentes distintos.

O primeiro componente, denominado componente de mercado, consiste em uma matriz de dimensão 21×22 , na qual a primeira dimensão corresponde a uma janela deslizante dos vinte e um pregões mais recentes, equivalente a aproximadamente um mês de negociação, e a segunda dimensão reúne as vinte e duas variáveis que descrevem o estado do mercado em cada um desses dias. Dessas variáveis, vinte derivam dos cinco ativos da carteira, contemplando, para cada ativo, os quatro indicadores detalhados na seção 3.3.3: o log-retorno diário, a volatilidade móvel e os índices de Sharpe e de Sortino calculados sobre a janela móvel. As duas variáveis restantes correspondem aos indicadores macroeconômicos, a taxa SELIC

diária e o IPCA, cuja inclusão constitui a principal contribuição do desenho do ambiente, ao conferir ao modelo a capacidade de contextualizar suas decisões de alocação no regime macroeconômico vigente. A adoção de uma janela temporal, em lugar de uma observação pontual, permite ao modelo inferir tendências e padrões de momentum a partir da trajetória recente das variáveis, em consonância com a estrutura sequencial inerente aos mercados financeiros.

O segundo componente, denominado componente de pesos, é um vetor de dimensão cinco que representa a alocação corrente da carteira entre os ativos no instante de decisão. A inclusão explícita dos pesos vigentes no espaço de estados é uma decisão de desenho deliberada e necessária: sem o conhecimento de sua posição atual, o modelo seria incapaz de avaliar a magnitude do rebalanceamento implicado por uma nova decisão de alocação e, por conseguinte, de antecipar os custos de transação dela decorrentes. Dessa forma, o estado integra simultaneamente a informação externa do mercado e a informação intrínseca da própria carteira, condição indispensável para que o modelo aprenda a ponderar o benefício de uma realocação contra seu custo operacional.

3.4.2 Espaço de Ações e Restrição de Concentração

O espaço de ações foi modelado como contínuo, em conformidade com a natureza do problema de alocação de portfólio, em que a decisão consiste em determinar, a cada instante, a fração do capital a ser alocada em cada um dos cinco ativos. Formalmente, o espaço de ações é definido como uma caixa retangular $A \subset \mathbb{R}^5$, com cada componente restrita ao intervalo $[-10, +10]$. As ações emitidas pelo Ator não correspondem diretamente aos pesos da carteira, mas a vetores de *logits* que são posteriormente transformados em uma distribuição de pesos válida por meio de uma função de mapeamento determinística. Essa separação entre o espaço bruto de saída da rede neural e o espaço efetivo de pesos é uma prática consolidada em aplicações de Aprendizado por Reforço a problemas de alocação contínua, pois permite ao modelo operar em um espaço euclidiano irrestrito durante a otimização, ao passo que as restrições do vetor de pesos, não negatividade e soma unitária, são impostas externamente pela função de mapeamento (JIANG; XU; LIANG, 2017).

A transformação dos *logits* em pesos é realizada em duas etapas. A primeira aplica a função *softmax*, que mapeia o vetor $\mathbb{R}^5$ para o vetor de probabilidades, garantindo simultaneamente a não negatividade dos pesos e a soma unitária:

$$w'_i = \frac{\exp(a_i - \max(\mathbf{a}))}{\sum_j \exp(a_j - \max(\mathbf{a}))}$$

A subtração do valor máximo no expoente é uma medida de estabilidade numérica: quando os *logits* assumem valores muito grandes, a operação $\exp(a_i)$ da fórmula clássica pode exceder os limites de representação de ponto flutuante e gerar erro computacional. A subtração mantém os argumentos da exponencial sempre não-positivos, evitando esse transbordamento sem alterar o resultado matemático, propriedade que decorre do cancelamento da constante subtraída entre numerador e denominador da fração. O vetor de pesos resultante dessa primeira etapa é válido no sentido de pertencer ao vetor de probabilidades, mas não impõe qualquer limite à concentração em um único ativo, motivando a segunda etapa do mapeamento.

A segunda etapa introduz uma restrição adicional, motivada por considerações de risco e justificada empiricamente no processo iterativo descrito na seção 3.5, onde é citado que nenhum ativo pode receber peso superior a 40% da carteira. Essa restrição materializa, no nível do mapeamento da ação, o princípio de diversificação mínima, impedindo que o modelo concentre excessivamente o portfólio em um único ativo, comportamento que, embora possa maximizar o retorno em períodos específicos, eleva substancialmente o risco idiossincrático.

A imposição do limite é realizada por meio de um algoritmo iterativo de redistribuição. A cada iteração, identificam-se os ativos cujo peso excede o limite máximo $W_{max} = 0,40$. O excedente é truncado para 0,40 e o total excedente, é redistribuído proporcionalmente aos demais ativos cujo peso é estritamente inferior ao limite, na proporção de seus pesos correntes. O procedimento é repetido até que nenhum peso exceda o limite, convergindo em poucas iterações dado que o peso máximo da carteira é não-crescente a cada passo. O resultado é um vetor de pesos no vetor de probabilidades que respeita simultaneamente a restrição de soma unitária e a restrição de concentração máxima.

3.4.3 Função de Recompensa

A função de recompensa é o componente do Processo de Decisão de Markov responsável por traduzir os objetivos do problema em um sinal numérico que orienta o aprendizado da política. Sua especificação constitui uma das decisões de desenho mais delicadas em aplicações de Aprendizado por Reforço, pois o modelo otimiza estritamente aquilo que recebe como recompensa. Qualquer desalinhamento entre o sinal escolhido e o objetivo subjacente do problema resulta em políticas que maximizam o critério literal em detrimento da intenção do pesquisador.

Neste trabalho, adotou-se como recompensa o log-retorno líquido do portfólio a cada pregão, isto é, o log-retorno bruto da carteira deduzido do custo logarítmico de transação incorrido na transição entre alocações:

$$R_t = \ln(1 - \tau_t \cdot f) + \langle w_t^{\text{exec}}, r_t \rangle$$

onde T_t representa o turnover executado no instante t , $f = 0,0003$ é a taxa proporcional de transação, que procura refletir a estrutura de custos efetivamente incidente sobre a negociação de fundos de índice no mercado à vista da B3 (B3, 2024), w_t^{exec} é o vetor de pesos efetivamente em vigor no pregão e r_t é o vetor de log-retornos dos ativos no mesmo instante. A modelagem detalhada do turnover, da taxa de transação e da condição que determina a execução ou não do rebalanceamento é apresentada na seção 3.4.4.

A definição do log-retorno líquido como sinal de recompensa repousa sobre um tripé de justificativas teóricas e práticas. Inicialmente, a própria matemática do log-retorno garante que a soma contínua das recompensas, objetivo de otimização central do algoritmo PPO, reflita com exatidão o crescimento acumulado do patrimônio, alinhando o incentivo do modelo à métrica real de sucesso do investidor (CAMPBELL; LO; MACKINLAY, 1997). Em segundo lugar, sob a ótica da aprendizagem, embutir o custo de transação diretamente no sinal de retorno assegura que o modelo sinta o ônus financeiro imediato de cada operação (JIANG; XU; LIANG, 2017). Essa mecânica dispensa a programação de regras rígidas e permite que o algoritmo internalize organicamente o dilema entre corrigir o desvio estratégico da carteira e preservar o capital contra o atrito gerado pelas taxas da bolsa.

A terceira justificativa é estritamente empírica e validada pelo método experimental. Conforme documentado na evolução dos lotes de treinamento (seção

3.5), testou-se inicialmente uma variante da função que subtraía a taxa Selic diária do retorno do portfólio, sob a premissa de que penalizar o modelo pelo custo de oportunidade da renda fixa induziria uma postura mais conservadora. Na prática, essa formulação prejudicou a convergência do modelo. A simplificação para o log-retorno líquido puro demonstrou ser significativamente mais estável, consolidando-se como a especificação definitiva.

Nesse sentido, cumpre registrar uma decisão arquitetural deliberada: a recompensa adotada não penaliza explicitamente métricas de risco, como o *Maximum Drawdown* ou a volatilidade. Embora estudos na fronteira do *Deep Reinforcement Learning* financeiro atestem a viabilidade de funções de recompensa ajustadas ao risco (ALMAHDI; YANG, 2017), este trabalho adotou uma abordagem minimalista com relação à modelagem da função objetivo. Optou-se por delegar a contenção de riscos à própria modelagem do ambiente como a trava de 40% por ativo, a banda de tolerância para rebalanceamento e os custos operacionais. Essa escolha garante que o comportamento defensivo do modelo seja uma adaptação natural às regras do mercado, isolando com clareza o efeito do ambiente sobre a política aprendida e reservando a exploração de recompensas complexas para trabalhos futuros.

3.4.4 Dinâmica de Transação e Custo de Rebalanceamento

A última peça do Processo de Decisão de Markov é a dinâmica que rege a transição entre alocações sucessivas e o custo associado a essa transição. Esta seção descreve os três elementos que compõem essa dinâmica: a métrica de turnover empregada para quantificar o desvio entre a alocação corrente e a alocação-alvo proposta pelo modelo, o critério de acionamento do rebalanceamento e a função de custo aplicada quando a operação é efetivamente executada.

A cada pregão, a política produz um vetor de pesos-alvo w_t^{alvo} resultado da transformação descrita na seção 3.4.2, ao qual se contrapõe o vetor de pesos w_t^{atual} efetivamente em vigor na carteira no início do período. A medida adotada para quantificar a magnitude do rebalanceamento implicado pela ação proposta é o maior desvio absoluto individual entre os dois vetores:

$$\tau_t = \max_i |w_{i,t}^{alvo} - w_{i,t}^{atual}|$$

Em termos operacionais, essa expressão calcula, para cada um dos cinco ativos da carteira, a diferença em módulo entre o peso-alvo proposto pela política e o peso efetivamente vigente, e adota como turnover o maior desses cinco valores. A escolha dessa métrica, em detrimento da soma dos desvios absolutos sobre todos os ativos, decorre da fundamentação teórica apresentada e do processo iterativo de desenvolvimento documentado na seção 3.5. Conforme demonstrado por Masters (2003) e corroborado empiricamente pela Vanguard Research (2019), a prática institucional de gestão de portfólio define bandas de tolerância em torno do peso-alvo de cada ativo individualmente, e o rebalanceamento é acionado quando qualquer ativo extrapola sua banda. Essa formulação, que considera o maior desvio entre os ativos e não a soma agregada dos pequenos desvios, protege a carteira contra concentrações específicas sem desperdiçar capital em ajustes desnecessários decorrentes de oscilações simultâneas e pequenas em múltiplos ativos.

O rebalanceamento é efetivamente acionado apenas quando o turnover supera um limiar predefinido τ_{min} , ou seja, se $\tau_{alvo} > \tau_{min}$, então $w_t^{exec} = w_t^{alvo}$. Caso contrário, $w_t^{exec} = w_t^{atual}$. Esse mecanismo materializa, computacionalmente, a zona de não-transação teorizada por Leland (1996) e formalizada por Masters (2003) no contexto de bandas de tolerância, região do espaço de alocações dentro da qual o custo de rebalancear supera o benefício de restabelecer a alocação-alvo, tornando a inação a decisão ótima. O valor de $\tau_{min} = 0,05$ (cinco pontos percentuais) foi adotado em consonância com a recomendação empírica da Vanguard Research (2019), segundo a qual uma banda de tolerância de cinco pontos percentuais representa o equilíbrio adequado entre o custo de transação cumulativo e a aderência à alocação-alvo. A validação empírica desse valor no contexto específico deste trabalho é discutida no processo iterativo descrito na seção 3.5.

Quando o rebalanceamento é acionado, o custo de transação é calculado como o produto entre o turnover executado e a taxa proporcional f , atuando sobre o valor corrente do portfólio. A taxa proporcional foi fixada em $f = 0,0003$, equivalente a 0,03% do volume rebalanceado, valor calibrado para refletir, de forma conservadora, os custos explícitos médios incidentes sobre operações com ETFs de alta liquidez negociados na B3. Conforme discutido na seção 2.5.1, este parâmetro é uma simplificação proposital da estrutura completa de custos de transação descrita por

Almgren e Chriss (2000), na qual coexistem componentes lineares e não lineares. A modelagem aqui adotada captura o componente linear, dominante em ETFs de alta liquidez e baixo volume relativo de transação, e abstrai os componentes não lineares de impacto de mercado, justificáveis apenas para volumes muito superiores aos contemplados pelo capital inicial fixado em R\$10.000,00.

A integração do custo de transação ao sinal de recompensa, conforme definido na seção 3.4.3, ocorre por meio do componente *cost_log*, que expressa o efeito do custo sobre o log-retorno da carteira no instante da operação:

$$\text{cost_log}_t = \ln(1 - \tau_t \cdot f)$$

Cumprir destacar que essa formulação preserva a propriedade aditiva do log-retorno: o log-retorno líquido do portfólio em um pregão de rebalanceamento corresponde à soma exata do log-retorno bruto da alocação executada e do log-custo da operação. Quando o turnover não supera o limiar de acionamento e o rebalanceamento não é efetivado, o custo é nulo e a alocação permanece inalterada, preservando inteiramente o log-retorno bruto. Essa estrutura incentiva o modelo a internalizar, ao longo do treinamento, o *trade-off* entre desvio estratégico e atrito operacional, sem que essa relação precise ser explicitamente codificada em qualquer outra parte do ambiente.

3.5 PROCESSO ITERATIVO DE REFINAMENTO DO MODELO

A configuração final do modelo, em termos de hiperparâmetros do algoritmo, função de recompensa e dinâmica do ambiente, não foi determinada *a priori* a partir de valores recomendados pela literatura ou por intuição metodológica. Em vez disso, foi construída por meio de um processo iterativo de refinamento empírico, organizado em cinco lotes sucessivos de experimentação, cada um introduzindo uma modificação pontual e controlada em relação ao anterior. Essa abordagem aproxima o desenvolvimento do modelo do método experimental clássico: cada lote constitui uma rodada de hipótese, intervenção e diagnóstico, na qual o efeito de uma alteração isolada é mensurado e interpretado antes que a próxima decisão de projeto seja tomada.

A motivação para essa estratégia decorre de uma característica do problema que não pode ser endereçada por escolhas teóricas isoladas a dinâmica do

treinamento de um modelo PPO sobre o ambiente de portfólio é simultaneamente sensível a múltiplos parâmetros, como taxa de aprendizado, tamanho do *rollout*, função de recompensa, dinâmica de transação e restrições estruturais, cujas interações não são plenamente previsíveis a partir da formulação matemática do algoritmo. Modificar simultaneamente vários desses parâmetros tornaria impossível atribuir mudanças no comportamento do modelo a uma causa específica, comprometendo o aprendizado metodológico do próprio processo. A organização em lotes garante, ao contrário, a rastreabilidade entre cada decisão de projeto e a evidência empírica que a fundamentou.

Cada lote é descrito a seguir com a mesma estrutura: a configuração específica adotada, os resultados quantitativos observados sobre o conjunto de dados utilizado para teste, o diagnóstico extraído desses resultados e a decisão de projeto que dele decorreu, materializando-se na configuração do lote seguinte. A última configuração resultante desse processo, o Lote 5, corresponde àquela efetivamente empregada na avaliação final dos resultados, descrita no capítulo 4. Os resultados aqui apresentados correspondem a médias e desvios-padrão obtidos sobre múltiplas execuções independentes (*seeds*) de cada configuração, em conformidade com o protocolo *multi-seed* discutido na seção 3.6, que visa controlar a variância intrínseca decorrente da inicialização aleatória dos parâmetros da rede neural.

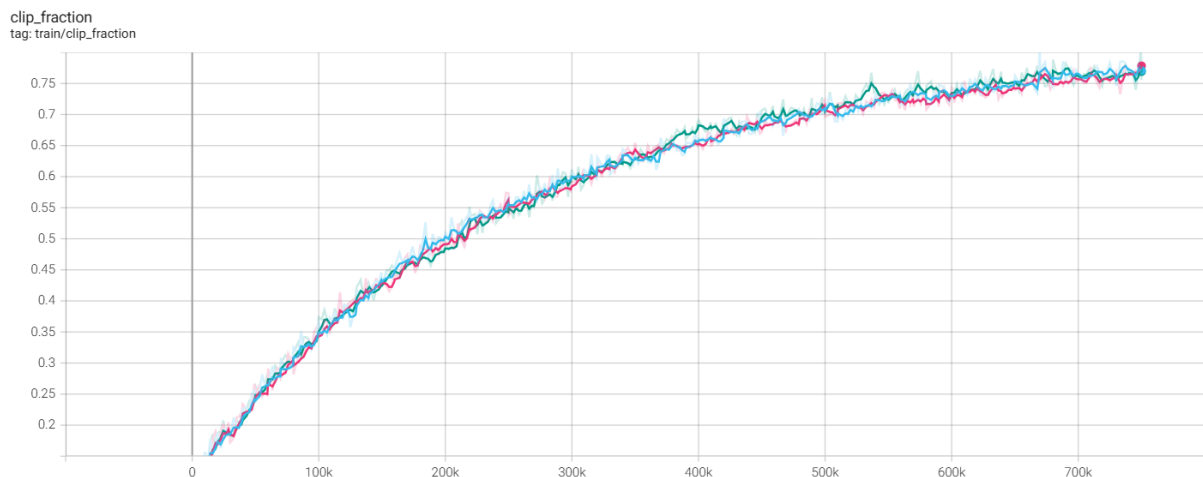
3.5.1 Configuração Inicial e Diagnóstico de Instabilidade

A primeira iteração de experimentação (lote 1) adotou uma configuração inicial baseada nos valores padrão do algoritmo PPO conforme implementado na biblioteca Stable-Baselines3 (RAFFIN et al., 2021), preservados sem modificações como ponto de partida deliberado para avaliar o comportamento do algoritmo sob a configuração de referência. A taxa de aprendizado foi fixada em 3×10^{-4} , valor padrão tipicamente empregado em arquiteturas baseadas em perceptron multicamadas, o tamanho do *rollout* foi definido em 2.048 transições por atualização e o orçamento total de treinamento estabelecido em 750.000 *timesteps*. A função de recompensa correspondeu ao log-retorno líquido do portfólio, conforme detalhado na seção 3.4.3, e o limiar de acionamento do rebalanceamento foi inicialmente parametrizado em 2%, sem qualquer restrição estrutural sobre a concentração máxima por ativo. Os demais

parâmetros, em particular o fator de desconto $\gamma = 0,99$ e o tamanho de minibatch igual a 64, foram mantidos em seus valores convencionais e preservados ao longo de todas as iterações subsequentes.

O treinamento foi conduzido sobre três sementes aleatórias independentes, conforme o protocolo adotado. A inspeção das métricas internas do algoritmo, registradas ao longo do treino e visualizadas por meio do framework TensorBoard, revelou um padrão consistente entre as três execuções que se mostraria determinante para o redirecionamento metodológico subsequente. As Figuras 2, 3 e 4 apresentam a trajetória de três métricas-chave de diagnóstico ao longo do treinamento dos três modelos. Em todas elas, o eixo das abscissas registra o número acumulado de passos de interação com o ambiente, estendendo-se de zero a aproximadamente setecentos e cinquenta mil transições, ao passo que o eixo das ordenadas registra o valor da métrica correspondente, adimensional em todos os três casos.

Figura 2 - Métricas internas de treinamento do Lote 1: clip_fraction



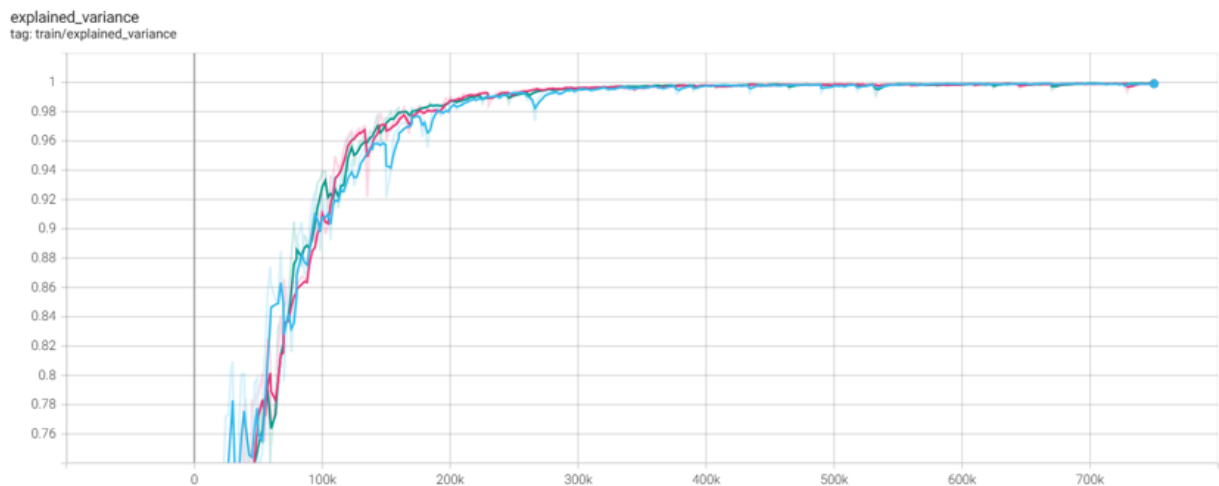
Fonte: Elaboração própria (2026)

Figura 3 - Métricas internas de treinamento do Lote 1: *approx_kl*



Fonte: Elaboração própria (2026)

Figura 4 - Métricas internas de treinamento do Lote 1: *explained_variance*



Fonte: Elaboração própria (2026)

A análise conjunta dessas trajetórias permite um diagnóstico preciso da dinâmica de treinamento. A *clip_fraction*, proporção de amostras em que a razão de probabilidade entre a nova política e a política anterior foi efetivamente truncada pelo mecanismo de clipping do PPO, apresentou crescimento substancial ao longo de todo o treinamento, partindo de aproximadamente 0,20 e atingindo o valor médio final de 0,77, com comportamento essencialmente idêntico entre as três sementes. O mecanismo de *clipping* introduzido por Schulman et al. (2017) tem como propósito explícito conter atualizações de política que se distanciem excessivamente da política anterior, mantendo o aprendizado dentro de uma região de confiança em torno da

distribuição de amostragem, uma proporção crescente e sustentadamente elevada de amostras clipadas indica, portanto, que o algoritmo opera sistematicamente no limite dessa restrição, sinalizando que cada atualização da política tenta efetuar passos excessivamente grandes em relação ao espaço de soluções vizinhas, condição que caracteriza regime de instabilidade.

A divergência KL aproximada (*approx_kl*) confirmou o mesmo diagnóstico por uma via independente. A magnitude média das atualizações de política cresceu de aproximadamente 0,01 no início do treinamento para 0,645 ao final, valor cerca de trinta vezes superior ao patamar de 0,02 observado em implementações estáveis de PPO (HUANG et al., 2022). Esse padrão indica que, à medida que o treinamento avançava, o algoritmo passou a propor atualizações progressivamente mais agressivas dos parâmetros da rede, refletindo a incapacidade da configuração corrente em estabilizar a política em torno de um ótimo local.

A *explained_variance* da função de valor, em contrapartida, evoluiu de forma exemplar, partindo de valores próximos de 0,75 nos primeiros pregões e estabilizando-se em 0,99 após aproximadamente 200.000 *timesteps*, patamar mantido até o final do treinamento. Esse resultado fornece uma evidência argumentativa importante: o Crítico (função de valor) aprendeu o ambiente com altíssima precisão, sendo capaz de estimar quase perfeitamente o retorno esperado a partir de cada estado. A instabilidade observada nas métricas anteriores, portanto, não decorre de uma falha do Crítico, mas reside especificamente no Ator, a política contínua que mapeia estados em pesos de portfólio. Essa distinção é metodologicamente importante, pois orientou as decisões subsequentes a atuar sobre parâmetros que regulam diretamente a dinâmica de atualização da política, e não sobre a estrutura da rede de valor.

A avaliação *out-of-sample* dos três modelos resultantes corroborou o diagnóstico de instabilidade. O retorno total médio sobre o conjunto de teste situou-se em 23,86%, com desvio-padrão entre sementes de 3,26 pontos percentuais, dispersão substancial considerando-se a magnitude do retorno. A volatilidade anualizada média foi de 10,85%, com Sharpe ligeiramente negativo (média -0,043), e o número médio de rebalanceamentos por episódio atingiu 413, gerando um custo operacional médio de R\$ 473,33 sobre o capital inicial de R\$ 10.000. Esses dois últimos indicadores, quantidade de rebalanceamentos e custo cumulativo, revelaram-

se particularmente informativos, sugerindo um comportamento de hiperatividade do modelo, que realizava ajustes frequentes na carteira, gerando atrito operacional desproporcional ao benefício marginal das realocações.

O diagnóstico extraído desta primeira iteração organizou-se em três frentes. Em primeiro lugar, a instabilidade do Ator, evidenciada pelo crescimento monotônico da *clip_fraction* e da *approx_kl*, apontava para a necessidade de atualizações mais conservadoras da política, o que motivou, na iteração seguinte, a redução da taxa de aprendizado e a ampliação do tamanho do *rollout*, ambos parâmetros que regulam diretamente a magnitude e a estabilidade dos passos de atualização. Em segundo lugar, a hiperatividade transacional sugeria que o limiar de 2% para o acionamento do rebalanceamento era demasiado permissivo, exigindo elevação para um valor mais compatível com a prática institucional documentada na literatura (seção 2.5). Em terceiro lugar, ainda que insuficiente isoladamente, a evidência justificava a extensão do orçamento de treinamento, dado que o crescimento progressivo das métricas de instabilidade ao longo de toda a janela de 750.000 *timesteps* indicava que o algoritmo não havia atingido um regime estacionário, comportamento que poderia tanto refletir incapacidade da configuração quanto orçamento computacional insuficiente. A configuração resultante dessas três decisões é apresentada na próxima subseção.

3.5.2 Ajuste de Hiperparâmetros e Emergência de Hiperatividade

A segunda iteração de experimentação (lote 2) atacou diretamente as três fontes de instabilidade identificadas no Lote 1, mediante intervenção combinada em parâmetros que regulam a dinâmica de atualização da política, a frequência de transações e o orçamento de treinamento. A taxa de aprendizado foi reduzida em um fator de três, passando de 3×10^{-4} para 1×10^{-4} , com o objetivo de tornar as atualizações de política mais conservadoras. O tamanho do *rollout* foi dobrado, de 2.048 para 4.096 transições, ampliando o volume de experiência sobre o qual cada atualização é calculada e, por consequência, reduzindo a variância do estimador de gradiente. O orçamento total de treinamento foi estendido para 1.000.000 de *timesteps*, oferecendo ao algoritmo margem adicional para convergência. Por fim, o limiar de acionamento do rebalanceamento foi elevado de 2% para 5%, em consonância com a recomendação institucional documentada na seção 2.5. A função

de recompensa e a ausência de restrição estrutural sobre a concentração por ativo foram mantidas idênticas ao Lote 1.

O treinamento foi conduzido sobre cinco sementes aleatórias independentes, ampliando-se o protocolo *multi-seed* em relação à iteração anterior. As métricas internas de treinamento revelaram melhoria substancial em relação ao diagnóstico de instabilidade do Lote 1. A *clip_fraction* média ao final do treinamento situou-se em 0,47, contra os 0,77 anteriores, redução de aproximadamente 40% que retirou o algoritmo do regime de instabilidade severa. A divergência KL aproximada apresentou redução ainda mais expressiva, caindo de 0,645 para 0,087, valor cerca de oito vezes menor que o do lote anterior e próximo da faixa típica documentada para problemas de controle contínuo bem-configurados. A *explained_variance* permaneceu em 0,99, confirmando a manutenção de um Crítico estável e bem ajustado ao ambiente.

Os resultados *out-of-sample* dos cinco modelos, contudo, revelaram um padrão inesperado e diagnosticamente importante. O retorno total médio elevou-se substancialmente em relação ao Lote 1, atingindo 31,87%, aumento absoluto de oito pontos percentuais. No entanto, esse ganho médio foi acompanhado por um grande aumento na dispersão entre sementes. O desvio-padrão do retorno entre os cinco modelos foi de 17,02 pontos percentuais, mais de cinco vezes superior ao do Lote 1. O retorno individual variou de 14,99% (seed 14) a 59,35% (seed 11), amplitude superior a quarenta pontos percentuais para a mesma configuração e o mesmo dataset, o que caracteriza o resultado como praticamente sorteio entre inicializações da rede neural. A volatilidade média subiu para 13,34%, o *drawdown* máximo médio aprofundou-se para -9,23%, e o custo operacional médio elevou-se a R\$ 696, indicando atrito transacional ainda mais pronunciado do que no Lote 1.

A interpretação conjunta desses resultados configura o diagnóstico central deste lote. A estabilização do treinamento, evidenciada pela queda das métricas de *clipping* e divergência KL, indicou que os ajustes de hiperparâmetros foram bem-sucedidos em conter a instabilidade da política durante o processo de aprendizado. Entretanto, a estabilidade do treinamento não se traduziu em estabilidade do desempenho. A ausência de qualquer restrição estrutural sobre a concentração de ativos permitiu que o modelo convergisse, em diferentes sementes, para políticas radicalmente distintas, algumas concentrando capital em ativos de maior retorno e maior risco, outras adotando alocações mais diversificadas. Em outras palavras, a

função de recompensa baseada exclusivamente no log-retorno líquido oferecia múltiplos ótimos locais aproximadamente equivalentes em valor esperado, mas correspondentes a perfis de risco substancialmente distintos, e a escolha entre eles dependia essencialmente da inicialização aleatória da rede.

O diagnóstico extraído desta iteração orientou as duas iterações seguintes a investigar mecanismos complementares para o controle do risco. A hipótese inicial, explorada no Lote 3, foi de natureza algorítmica: modificar a função de recompensa de modo a penalizar diretamente o modelo pela ausência de superação da taxa livre de risco, internalizando o custo de oportunidade no próprio sinal de aprendizado. A hipótese subsequente, materializada no Lote 4, foi de natureza estrutural: impor uma restrição rígida sobre a concentração máxima por ativo, removendo do espaço de soluções as políticas excessivamente concentradas independentemente de seu valor esperado de retorno.

3.5.3 Investigação de Função de Recompensa Penalizada pela Taxa Livre de Risco

A terceira iteração (lote 3) investigou uma hipótese de natureza algorítmica para o problema de dispersão entre sementes diagnosticado no Lote 2. A motivação teórica derivou-se da observação de que, em períodos de taxa básica de juros elevada, o custo de oportunidade de manter exposição a ativos de risco é particularmente alto: cada unidade de risco assumida pelo modelo deveria ser compensada não apenas por um retorno nominal positivo, mas por um retorno superior àquele obtido sem qualquer risco em renda fixa. A função de recompensa adotada nos lotes anteriores, log-retorno líquido, não internalizava essa exigência. Ela recompensava o modelo por qualquer retorno positivo, ainda que inferior à taxa livre de risco. A hipótese, portanto, foi de que substituir o sinal de recompensa por uma versão penalizada pela SELIC diária induziria o modelo a desenvolver políticas mais defensivas em regimes de juros elevados, eventualmente reduzindo a dispersão entre sementes ao restringir o espaço de ótimos locais aceitáveis.

A função de recompensa modificada foi formalmente definida como:

$$R'_t = R_t - r_{f,t} = \ln(1 - \tau_t \cdot f) + \langle w_t^{\text{exec}}, r_t \rangle - r_{f,t}$$

onde $r_{f,t}$ representa o log-retorno diário da SELIC no instante t . Todos os demais elementos da configuração, taxa de aprendizado, tamanho de *rollout*, limiar de rebalanceamento, ausência de restrição de concentração, foram mantidos rigorosamente idênticos ao Lote 2, isolando deliberadamente o efeito da modificação na função de recompensa. O treinamento foi conduzido sobre três sementes independentes.

As métricas internas de treinamento foram praticamente idênticas às do Lote 2: *clip_fraction* média final de 0,47, *approx_kl* final de 0,08 e *explained_variance* estabilizada em 0,99. Essa invariância das métricas de treino confirma, do ponto de vista experimental, que a modificação isolada na função de recompensa não alterou a dinâmica de aprendizado do PPO, o algoritmo treinou com a mesma estabilidade, e qualquer diferença observada no desempenho *out-of-sample* pode ser atribuída exclusivamente à mudança no sinal de recompensa, e não a uma variação no processo de otimização.

Os resultados *out-of-sample*, no entanto, refutaram a hipótese inicial. O retorno total médio caiu de 31,87% no Lote 2 para 23,41%, uma redução de aproximadamente oito pontos percentuais. O índice de Sharpe, que havia se tornado positivo no Lote 2, voltou ao terreno negativo, com média de -0,049. O *drawdown* máximo médio apresentou melhoria apenas marginal, de -9,23% para -7,92%, insuficiente para compensar a perda de retorno. A dispersão entre sementes manteve-se elevada, com desvio-padrão de 6,58 pontos percentuais no retorno, o que indica que o objetivo declarado da intervenção, homogeneizar o comportamento entre inicializações por meio da restrição do espaço de ótimos locais, também não foi alcançado.

A interpretação extraída desta iteração foi clara. A subtração da SELIC do sinal de recompensa introduziu um termo negativo persistente no aprendizado, particularmente penalizante em regimes de juros elevados, exatamente o cenário em que se concentra o período de teste deste trabalho. Em vez de orientar o modelo a buscar alocações que efetivamente superassem a taxa livre de risco, a recompensa modificada parece ter induzido políticas conservadoras subótimas: o modelo, ao ser sistematicamente penalizado por permanecer exposto ao mercado, aprendeu a evitar exposição mesmo em situações nas quais o retorno marginal de alguns ativos justificaria-a. Em outras palavras, a penalização operou como um viés de aversão a

risco sem fornecer ao modelo, simultaneamente, sinal informativo suficiente para discriminar entre oportunidades de risco-retorno favoráveis e desfavoráveis.

A constatação metodológica decorrente foi de que o controle do risco não poderia ser implementado eficazmente por meio de penalização aditiva no sinal de recompensa, ao menos não na forma simples aqui investigada. A hipótese da iteração seguinte, materializada no Lote 4, deslocou-se para uma alteração do ambiente. Em vez de tentar induzir comportamento defensivo por meio do sinal de aprendizado, optou-se por restringir diretamente o espaço de ações disponíveis ao modelo, impedindo a formação de carteiras excessivamente concentradas mediante a imposição de um limite máximo de alocação por ativo.

3.5.3 Restrição Estrutural de Concentração e o Surgimento do Atrito Operacional

A quarta iteração (lote 4) abandonou a abordagem algorítmica testada no Lote 3 em favor de uma intervenção estrutural sobre o espaço de ações do modelo. A hipótese norteadora foi de que a alta dispersão entre sementes observada no Lote 2 decorria, fundamentalmente, da existência de múltiplos ótimos locais aproximadamente equivalentes em retorno esperado, mas correspondentes a perfis de concentração radicalmente distintos. Em vez de tentar induzir comportamento defensivo por meio do sinal de recompensa, optou-se por remover diretamente do espaço de soluções as políticas excessivamente concentradas, impondo uma restrição rígida sobre o peso máximo de qualquer ativo individual.

A intervenção materializou-se na introdução do limite $w_{max} = 0,40$ aplicado durante a transformação dos *logits* em pesos, conforme detalhado na seção 3.4.2, mediante o algoritmo iterativo de redistribuição ali descrito. Nenhum ativo poderia, portanto, receber mais de 40% do capital total da carteira, independentemente do que a política da rede neural recomendasse. A função de recompensa retornou à formulação simples do Lote 2 (log-retorno líquido, sem penalização pela SELIC), em vista do diagnóstico negativo da iteração anterior. Os demais parâmetros foram mantidos, e o treinamento foi conduzido sobre cinco sementes independentes.

As métricas internas de treinamento apresentaram nova melhoria incremental em relação aos lotes precedentes: a *clip_fraction* média final reduziu-se a 0,44, a

approx_kl a 0,062 e a *explained_variance* manteve-se em 0,99. Embora as variações em relação ao Lote 2 sejam modestas, sua direção é consistente e admite interpretação teórica plausível: a restrição estrutural sobre o espaço de ações reduz o conjunto de políticas que o algoritmo precisa explorar, o que tende a diminuir, marginalmente, a magnitude das oscilações durante a otimização.

Os resultados *out-of-sample* confirmaram, em grande medida, a eficácia da intervenção quanto ao controle de risco, mas revelaram um efeito colateral inesperado que se mostraria determinante para a iteração final. A volatilidade média da carteira caiu de 13,34% no Lote 2 para 7,91%, redução de quase 40% que materializa o efeito direto da restrição de concentração sobre a exposição ao risco. O desvio-padrão do retorno entre sementes recuou de 17,02 para 3,65 pontos percentuais, indicando convergência substancialmente mais homogênea entre inicializações distintas. O *drawdown* máximo médio melhorou para -5,99%, contra -9,23% do Lote 2. Esses três indicadores, em conjunto, validam a hipótese estrutural de que a restrição de 40% efetivamente conteve a formação de carteiras excessivamente arriscadas e homogeneizou o comportamento entre sementes.

O retorno total médio, contudo, recuou para 20,36%, abaixo dos lotes anteriores. Mais relevante ainda, dois indicadores apontaram para um fenômeno operacional inesperado: o número médio de rebalanceamentos por episódio elevou-se a 445, superior ao do Lote 2 (381) e o mais alto observado até então; o custo cumulativo de transação atingiu R\$ 478 sobre o capital inicial de R\$ 10.000. Essa combinação, risco controlado mas custo operacional elevado, caracterizou um padrão de hiperatividade transacional sob restrição: o modelo, ao operar dentro de um espaço de ações limitado pela restrição de 40%, passou a executar ajustes pequenos e contínuos com frequência elevada, gerando atrito operacional desproporcional ao benefício marginal das realocações.

O diagnóstico extraído foi de natureza eminentemente técnica e remeteu à implementação interna do ambiente. A métrica utilizada para acionar o rebalanceamento, a soma dos desvios absolutos entre os pesos-alvo e os pesos correntes, conforme implementação inicial, mostrou-se demasiado sensível ao tipo específico de oscilação induzido pela restrição de concentração. Quando a política propunha, em pregões sucessivos, pequenas redistribuições de pesos entre múltiplos ativos para manter cada um dentro do limite de 40%, a soma desses pequenos

desvios podia facilmente ultrapassar o limiar de 5%, acionando o rebalanceamento ainda que nenhum ativo individualmente apresentasse desvio significativo em relação ao alvo. A próxima iteração investigou, portanto, uma modificação na métrica de turnover, alinhando-a com a prática institucional de bandas de tolerância individuais.

3.5.5 Refinamento da Métrica de Turnover e Configuração Final

A quinta e última iteração (lote 5) materializou a hipótese diagnóstica formulada ao final do Lote 4. A análise daquele lote havia identificado que a hiperatividade transacional sob restrição decorria não do comportamento intrínseco do modelo, mas da inadequação da métrica empregada para acionar o rebalanceamento ao tipo específico de oscilação induzido pela restrição estrutural de 40%. A correção, portanto, não exigia modificação de hiperparâmetros, da função de recompensa ou da arquitetura do modelo, mas tão somente da operação matemática que computa o turnover entre a alocação corrente e a alocação-alvo.

A modificação consistiu em substituir, na implementação do ambiente, o cálculo do turnover via soma dos desvios absolutos pelo cálculo via maior desvio absoluto individual, conforme já formalizado na seção 3.4.4:

$$\tau_t = \max_i |w_{i,t}^{\text{alvo}} - w_{i,t}^{\text{atual}}| \quad \text{em substituição à} \quad \tau_t = \sum_i |w_{i,t}^{\text{alvo}} - w_{i,t}^{\text{atual}}|$$

Trata-se, do ponto de vista da implementação, de uma alteração de uma única linha de código, a substituição da função `np.sum` por `np.max` na expressão de turnover. Do ponto de vista metodológico, contudo, a modificação alinha o critério de acionamento do rebalanceamento à prática institucional de bandas de tolerância individuais discutida na seção 2.6, segundo a qual o ajuste da carteira é acionado quando *qualquer* ativo extrapola sua banda, e não quando a soma agregada dos pequenos desvios ultrapassa um limite arbitrário. Todos os demais elementos da configuração foram mantidos rigorosamente idênticos ao Lote 4, garantindo que qualquer efeito observado pudesse ser atribuído exclusivamente à modificação cirúrgica na métrica de turnover. O treinamento foi conduzido sobre três sementes independentes.

As métricas internas de treinamento permaneceram essencialmente estáveis em relação ao Lote 4: *clip_fraction* média final de 0,44, *approx_kl* de 0,068 e *explained_variance* de 0,99. Essa invariância é coerente com a natureza da

modificação: a alteração na métrica de turnover não afeta a dinâmica de atualização da política propriamente dita, mas apenas o critério a posteriori que decide se uma transição entre alocações será efetivamente executada.

Os resultados *out-of-sample*, em contrapartida, revelaram melhoria substancial em múltiplas dimensões simultaneamente. O retorno total médio elevou-se de 20,36% no Lote 4 para 24,82%, incremento de aproximadamente 22% sobre o lote anterior. O Sharpe Ratio médio, que havia se mantido em terreno negativo durante todas as iterações anteriores, tornou-se marginalmente positivo (0,013) pela primeira vez no processo de desenvolvimento. O *drawdown* máximo médio recuou para -5,33%, e a volatilidade manteve-se essencialmente inalterada em 7,85%, preservando o efeito de controle de risco conquistado no Lote 4. A dispersão entre sementes situou-se em 4,01 pontos percentuais, em patamar comparável ao do Lote 4 e substancialmente inferior ao caos observado no Lote 2.

O efeito mais expressivo, contudo, manifestou-se nos indicadores operacionais. O número médio de rebalanceamentos caiu de 445 para 386, redução de 13%, e o custo cumulativo de transação reduziu-se de R\$478 para R\$211, queda de aproximadamente 56% sobre o capital inicial. Esses dois resultados, em conjunto, confirmam o diagnóstico do Lote 4. A hiperatividade transacional não derivava de uma necessidade legítima do modelo em ajustar a carteira, mas do acionamento errôneo do mecanismo de rebalanceamento por pequenos desvios distribuídos em múltiplos ativos. Ao adotar uma métrica que requer desvio individual significativo para acionar a operação, o ambiente passou a executar apenas os rebalanceamentos efetivamente necessários, eliminando o atrito desnecessário sem comprometer a aderência da carteira aos pesos-alvo da política.

A interpretação conjunta desses resultados, incremento de retorno, manutenção do controle de risco e redução expressiva de custo operacional, caracteriza o Lote 5 como a configuração que conjuga, de forma equilibrada, os ganhos individualmente identificados em cada uma das iterações anteriores: a estabilidade de treinamento conquistada no Lote 2, o controle de risco materializado no Lote 4 e a eficiência operacional emergente do refinamento aqui descrito. Essa configuração foi, portanto, adotada como a especificação final do modelo, sobre a qual se conduziu a avaliação multi-seed em larga escala apresentada no capítulo 4. Os hiperparâmetros e demais parâmetros do ambiente correspondentes a essa

configuração final encontram-se consolidados na Tabela descritiva de configuração final da seção 3.6.

3.6 CONFIGURAÇÃO FINAL E PROTOCOLO MULTI-SEED

Após a especificação detalhada do ambiente de simulação na seção 3.4 e a apresentação do processo iterativo de refinamento dos hiperparâmetros na seção 3.5, esta seção consolida a configuração efetivamente empregada nos experimentos e descreve o protocolo metodológico sob o qual a avaliação foi conduzida. A apresentação organiza-se em três aspectos complementares: a configuração consolidada do modelo e do ambiente; o protocolo multi-seed adotado para controlar a aleatoriedade intrínseca ao treinamento de redes neurais; e a forma estatística sob a qual os resultados são reportados no capítulo seguinte.

3.6.1 Configuração final

A configuração final do modelo, resultante do processo iterativo descrito na seção anterior, é apresentada de forma consolidada na Tabela 4. Essa configuração foi empregada em todas as avaliações reportadas no capítulo 4, abrangendo o período principal de teste e os cenários adicionais de estresse macroeconômico descritos na seção 3.7.

Tabela 4 - Configuração final do modelo

Parâmetro	Valor	Componente
Algoritmo	PPO	Modelo
Política	MultiInputPolicy	Modelo
Taxa de aprendizado	1×10^{-4}	Treinamento
Tamanho do rollout (N_STEPS)	4096	Treinamento
Tamanho de minibatch	64	Treinamento
Fator de desconto (γ)	0,99	Treinamento
Timesteps por treino	1.000.000	Treinamento
Frequência de avaliação	10.000 timesteps	Ambiente
Janela de observação	21 dias	Ambiente
Capital inicial	R\$ 10.000,00	Ambiente
Limiar de rebalanceamento	5%	Ambiente
Concentração máxima por ativo	40%	Ambiente
Função de recompensa	Log retorno líquido	Ambiente

Fonte: Elaboração própria (2026)

A justificativa para cada valor encontra-se distribuída ao longo do capítulo: os parâmetros do otimizador derivam diretamente do processo iterativo descrito na seção 3.5 (em particular, dos Lotes 1 e 2, que estabilizaram a dinâmica de treinamento); os parâmetros do ambiente, janela de observação, capital inicial e função de recompensa, foram detalhados nas seções 3.3.3 e 3.4; a taxa de transação reflete os custos institucionais discutidos na seção 2.5.1; e os parâmetros relativos ao rebalanceamento (limiar, métrica e concentração máxima) resultam, respectivamente, dos Lotes 2, 5 e 4 da seção 3.5.

3.6.2 Controle de Estocasticidade e Protocolo Multi-Seed

O treinamento de modelo de Aprendizado por Reforço apresenta sensibilidade documentada à inicialização aleatória dos parâmetros da rede neural, fenômeno reportado de forma consistente na literatura (HENDERSON et al., 2018). Em consequência, o desempenho avaliado a partir de uma única execução não constitui evidência confiável da qualidade da configuração adotada, podendo refletir tanto a competência da política aprendida quanto a sorte da inicialização. Para mitigar essa fonte de variabilidade, adotou-se um protocolo de execução *multi-seed*: cada cenário de avaliação foi submetido a múltiplas execuções independentes do procedimento de treinamento, cada qual partindo de uma semente aleatória distinta. Os resultados reportados no capítulo 4 correspondem, portanto, a estatísticas agregadas (média, mediana, desvio-padrão) sobre o conjunto de modelos resultantes, e não a desempenhos de execuções individuais.

A reprodutibilidade do procedimento e o controle integral das fontes de estocasticidade são assegurados pela fixação de uma semente aleatória por execução. Essa fixação é realizada por meio do parâmetro *seed* do construtor do algoritmo PPO da biblioteca Stable-Baselines3 (RAFFIN et al., 2021), que a propaga internamente aos geradores de números pseudoaleatórios da biblioteca PyTorch, que governa a inicialização dos pesos da rede neural e a amostragem da distribuição da política. A centralização do controle nesse gerador garante, de um lado, que execuções com a mesma semente sejam integralmente reproduzíveis e, de outro, que a variabilidade observada entre execuções com sementes distintas decorra exclusivamente da inicialização aleatória, pressuposto que fundamenta a interpretação dos resultados agregados pelo protocolo multi-seed.

3.6.2 Apresentação Estatística dos Resultados

Em decorrência do protocolo adotado, os resultados do capítulo 4 são reportados sob forma estatística, média e desvio-padrão sobre o conjunto de execuções, complementados pela mediana quando se justifica como descritor mais representativo da tendência central. A mediana é particularmente informativa em amostras nas quais sementes ocasionalmente convergem para políticas degeneradas, situações em que o modelo, por convergência prematura ou aprendizagem incompleta, adota comportamento essencialmente passivo. Tais ocorrências, embora minoritárias, deslocam a média para baixo de forma desproporcional, e o reporte conjunto de média e mediana oferece ao leitor uma caracterização mais fidedigna da distribuição de desempenhos.

À caracterização descritiva acrescenta-se, tanto no cenário principal quanto nos cenários de estresse, uma análise inferencial destinada a verificar se o desempenho do modelo difere de forma estatisticamente significativa das estratégias de referência. Dado que cada baseline é determinístico sobre o conjunto de teste, produzindo um único valor de retorno, e não uma distribuição, adotou-se o teste t de uma amostra, que confronta a média da distribuição de retornos do modelo, obtida sobre o conjunto de execuções, contra o valor fixo de cada estratégia de referência. Para cada comparação, testou-se a hipótese nula de igualdade entre a média do modelo e o valor do baseline, contra a hipótese alternativa bilateral de diferença, ao nível de significância de 5% ($\alpha = 0,05$); reportam-se, para cada par, a estatística t, o p-valor e o intervalo de confiança de 95% para a diferença de médias. O teste pressupõe a normalidade aproximada da distribuição de retornos entre execuções: no cenário principal, o tamanho amostral ($n = 201$) confere robustez a essa aproximação. Já nos cenários de estresse, conduzidos sobre amostra reduzida ($n = 30$), os resultados inferenciais são interpretados como evidência indicativa, em conjunto com a análise descritiva da distribuição, conforme detalhado nas seções correspondentes do capítulo 4.

3.7 PERÍODOS DE AVALIAÇÃO E CENÁRIOS DE TESTE

A validade externa de um modelo de Aprendizado por Reforço aplicado ao mercado financeiro depende, fundamentalmente, da heterogeneidade dos regimes

sobre os quais sua política é avaliada. A simples mensuração do desempenho em um único período histórico, ainda que testada e validada com conjuntos de dados diferentes dos usados para o treinamento, oferece evidência limitada da capacidade de generalização do modelo. A política pode ter aprendido a explorar regularidades específicas daquele regime, em detrimento de propriedades transferíveis a contextos distintos. Para mitigar esse risco e endereçar a hipótese central deste trabalho, que a incorporação de variáveis macroeconômicas ao espaço de estados confere ao modelo capacidade de adaptação entre regimes, o protocolo de avaliação foi estruturado em três cenários distintos, cada um correspondendo a uma janela temporal com características macroeconômicas particulares.

O primeiro e principal cenário de avaliação corresponde ao biênio 2023–2024, intervalo caracterizado por taxa básica de juros em patamar historicamente elevado e em trajetória de moderação gradual. O conjunto de treinamento abrange observações anteriores a janeiro de 2022; o conjunto de validação compreende o ano de 2022, sendo utilizado pelo *EvalCallback* para a seleção do melhor modelo durante o treinamento; e o conjunto de teste corresponde ao período de janeiro de 2023 a dezembro de 2024. Esse cenário foi adotado como referência principal por duas razões. Em primeiro lugar, oferece a janela mais recente disponível para avaliação, garantindo relevância para o contexto contemporâneo do investidor brasileiro. Em segundo, o regime de juros elevados configura um teste particularmente exigente: o retorno da taxa livre de risco torna-se concorrente direto da renda variável, e qualquer alocação que assuma risco precisa ser justificada por retorno marginal substancial, condição que distingue políticas tecnicamente competentes de políticas meramente sortudas.

O segundo cenário corresponde ao ano de 2020, marcado pelo choque sistêmico da pandemia de COVID-19 sobre os mercados financeiros globais. O treinamento abrange observações anteriores a janeiro de 2019, com validação sobre 2019 e teste sobre o intervalo de janeiro de 2020 a dezembro de 2020. Esse cenário foi selecionado por configurar um regime de estresse agudo de natureza distinta do cenário principal: a queda abrupta dos mercados de risco no primeiro trimestre de 2020 foi seguida de recuperação igualmente brusca, em meio a um ambiente macroeconômico de taxa básica de juros próxima de mínimas históricas. Avaliar o

modelo nesse regime permite investigar sua capacidade de resposta a choques súbitos, fenômeno qualitativamente distinto da dinâmica gradual do cenário principal.

O terceiro cenário corresponde ao biênio 2021–2022, intervalo no qual a taxa básica de juros brasileira experimentou um dos mais agressivos ciclos de elevação de sua história recente, partindo de 2% no início de 2021 e atingindo 13,75% ao final de 2022. O treinamento abrange observações anteriores a janeiro de 2020, com validação sobre 2020 e teste sobre o intervalo de janeiro de 2021 a dezembro de 2022. A justificativa para a inclusão desse cenário reside na natureza específica do desafio que ele impõe ao modelo: a elevação contínua dos juros altera substancialmente o cálculo de risco-retorno dos ativos da carteira ao longo do próprio período de avaliação, deteriorando os ativos de duração mais longa (em particular o IRFM11, indexado a títulos prefixados) e simultaneamente elevando a atratividade da renda fixa indexada à inflação. Trata-se, portanto, de um regime no qual a informação macroeconômica disponível ao modelo, em particular a trajetória da SELIC e do IPCA, é especialmente relevante para a tomada de decisão de alocação.

Em virtude da sensibilidade do PPO à inicialização aleatória discutida na seção 3.6, cada cenário foi avaliado sobre uma amostra de execuções independentes. O cenário principal foi avaliado sobre uma amostra alvo de 200 execuções, dimensionada para conferir poder estatístico adequado à comparação inferencial com os baselines de referência. Os cenários de estresse, COVID e expansão da SELIC, foram avaliados sobre trinta execuções cada, número que, embora inferior ao do cenário principal, é suficiente para a caracterização da tendência central e da dispersão do desempenho nesses regimes. A escolha por amostragem assimétrica é declarada deliberadamente: o cenário principal constitui o objeto central da avaliação quantitativa, ao passo que os cenários de estresse cumprem função complementar de investigação de robustez.

Cumprir destacar que, em todos os três cenários, o protocolo de treinamento e avaliação foi rigorosamente preservado. A configuração do modelo, como hiperparâmetros, função de recompensa, dinâmica do ambiente e procedimento multi-seed, é a mesma em todos os casos, alterando-se exclusivamente o intervalo temporal correspondente aos conjuntos de treinamento, validação e teste. Essa uniformidade metodológica garante que as diferenças observadas entre cenários sejam atribuíveis

às características intrínsecas de cada regime, e não a variações no procedimento experimental.

3.8 ESTRATÉGIA DE AVALIAÇÃO E BASELINES

A última etapa do percurso metodológico consiste na definição do protocolo de avaliação empregado para mensurar o desempenho do modelo treinado e confrontá-lo com estratégias de referência. Esta seção formaliza dois componentes desse protocolo: as estratégias *baseline* contra as quais o modelo é comparado, e o conjunto de métricas de desempenho empregadas na quantificação dos resultados.

A escolha das estratégias *baseline* foi orientada por dois critérios complementares. Em primeiro lugar, cada *baseline* deve representar uma abordagem distinta de gestão de portfólio, de modo que o conjunto delas cubra um espectro razoável de filosofias de investimento, desde a alocação inteiramente passiva até a alocação fundamentada em mensuração explícita de risco. Em segundo lugar, todos os *baselines* devem ser executados sobre o mesmo conjunto de teste *out-of-sample* e sob o mesmo intervalo temporal a que o modelo é submetido, garantindo que a comparação seja conduzida em condições idênticas.

A primeira estratégia de referência corresponde ao modelo mais simples de gestão passiva: aloca-se, no primeiro pregão do período de teste, frações iguais do capital entre os cinco ativos da carteira (20% cada), aplicando-se uma única vez o custo de transação de entrada, e mantém-se essa alocação inalterada até o final do período. Não há rebalanceamento subsequente: a carteira evolui livremente segundo os retornos compostos de cada ativo. Conforme discutido na seção 2.5, o B&H é a estratégia de referência canônica em estudos de gestão de portfólio justamente por não incorporar qualquer decisão ativa após a alocação inicial, qualquer estratégia ativa precisa demonstrar valor agregado *sobre* essa linha de base para justificar a complexidade adicional que introduz (ANG, 2014).

A segunda estratégia de referência consiste no investimento integral do capital à taxa básica de juros da economia brasileira, capitalizada diariamente. Em termos formais, o valor do portfólio em cada pregão t é dado por:

$$V_t^{\text{SELIC}} = V_0 \cdot \exp \left(\sum_{k=1}^t r_{f,k} \right)$$

onde $r_{f,k}$ é o log-retorno diário da SELIC no instante k e V_0 é o capital inicial. Essa baseline cumpre função teórica distinta da anterior: representa o retorno obtido sem assumir qualquer risco de mercado, constituindo o piso inferior racional para qualquer estratégia que pretenda justificar exposição a ativos de risco. Em termos das métricas de avaliação, a SELIC apresentará volatilidade próxima de zero e índices de Sharpe e Sortino numericamente extremos, dada a ausência de risco no denominador, característica que limita a comparabilidade direta dessas métricas, mas não compromete sua função como referência de retorno mínimo aceitável.

A terceira estratégia de referência corresponde a uma implementação clássica de paridade de risco, em que os pesos dos ativos são determinados de forma inversamente proporcional às suas volatilidades estimadas. A cada início de mês, a alocação é recalculada como:

$$w_{i,t} = \frac{\frac{1}{\sigma_{i,t}}}{\sum_j \frac{1}{\sigma_{j,t}}}$$

onde $\sigma_{i,t}$ é a volatilidade móvel do ativo i calculada conforme a seção 3.3.3. A rotação mensal dos pesos incorre em custo de transação proporcional ao turnover correspondente, calculado pela soma dos desvios absolutos entre os pesos novos e as correntes, com a mesma taxa de 0,03% aplicada ao modelo. A inclusão dessa baseline atende à motivação teórica de comparar o modelo não apenas contra estratégias passivas (B&H e SELIC), mas também contra uma estratégia ativa fundamentada em uma regra explícita de gestão de risco, permitindo verificar se o aprendizado por reforço agrega valor sobre regras determinísticas de alocação.

O desempenho do modelo e dos *baselines* é caracterizado por sete métricas, articuladas em três dimensões complementares: retorno, risco e relação risco-retorno. As três primeiras quantificam diretamente os ganhos: o retorno total acumulado sobre o período de teste, o retorno anualizado correspondente, e a volatilidade anualizada dos retornos diários. As duas seguintes quantificam dimensões específicas de risco: o drawdown máximo, conforme definido na seção 2.5.3, captura a maior queda percentual do patrimônio entre um pico e um vale subsequente; o número de rebalanceamentos e o custo cumulativo de transação quantificam o atrito operacional incorrido pela estratégia. As duas últimas sintetizam a relação entre retorno e risco: o índice de Sharpe e o índice de Sortino, ambos calculados de forma anualizada sobre

os retornos do período de teste, não as versões móveis utilizadas como variáveis de estado na seção 3.3.3, mas as formulações clássicas aplicadas à trajetória completa do portfólio. Complementarmente, o índice de Calmar, definido como a razão entre retorno anualizado e o módulo do drawdown máximo, oferece uma medida de eficiência por unidade de risco de cauda.

A taxa livre de risco empregada no cálculo dos índices de Sharpe e Sortino corresponde à média da SELIC diária observada no próprio período de teste, escolha que reflete o custo de oportunidade efetivo do investidor durante o intervalo de avaliação, preservando a comparabilidade dos índices entre os diferentes cenários discutidos na seção 3.7.

4 RESULTADOS

Os capítulos anteriores estabeleceram a base conceitual e metodológica que sustenta o presente trabalho. Este capítulo apresenta os resultados experimentais derivados da aplicação do modelo PPO especificado no capítulo 3, sob os três cenários de avaliação previamente definidos. A análise prioriza tanto a caracterização quantitativa do desempenho do modelo, em termos de retorno, risco e eficiência operacional, quanto sua confrontação direta com as estratégias de referência institucionais, com vistas a submeter à verificação empírica as hipóteses formuladas na introdução deste trabalho.

4.1 VISÃO GERAL

A análise inicia-se pelo cenário principal de avaliação, os anos de 2023-2024, objeto de exame quantitativo detalhado em virtude da amostra ampla de duzentas e uma execuções independentes do procedimento de treinamento. As estatísticas descritivas, os testes inferenciais sobre o desempenho do modelo em relação às estratégias de referência, a distribuição dos retornos entre sementes e os indicadores de comportamento operacional do modelo são apresentados de forma sistematizada nas subseções da seção 4.2, encerradas por uma síntese conclusiva do cenário.

Os cenários complementares de estresse macroeconômico, a pandemia de COVID-19 em 2020 e o ciclo de aperto monetário do biênio 2021-2022, são abordados, respectivamente, nas seções 4.3 e 4.4, sob estrutura análoga, porém em escala amostral reduzida, conforme protocolo definido na seção 3.7. Em ambos os casos, a análise prioriza a caracterização do desempenho do modelo sob regimes macroeconômicos qualitativamente distintos do regime principal e o contraste com as estratégias de referência aplicadas ao mesmo período.

A seção 4.5 consolida as evidências em uma análise comparativa entre os três regimes, com vistas a examinar a hipótese central deste trabalho à luz do conjunto dos resultados, a saber, a robustez do desempenho do modelo diante de contextos macroeconômicos distintos. O capítulo encerra-se na seção 4.6 com a discussão integrada dos achados, incluindo o reconhecimento explícito das limitações observadas no desempenho do modelo e suas implicações para trabalhos futuros.

Todos os resultados aqui apresentados referem-se exclusivamente a períodos *out-of-sample*, isto é, intervalos temporais não acessíveis ao modelo durante o procedimento de treinamento, conforme protocolo metodológico detalhado nas seções 3.6 e 3.7. As estatísticas reportadas, salvo menção em contrário, correspondem às agregações estatísticas (média, mediana, desvio-padrão) sobre o conjunto de modelos resultantes do protocolo multi-seed descrito na seção 3.6.2.

4.2 CENÁRIO PRINCIPAL

O cenário principal de avaliação compreende o biênio de 2023 a 2024, período em que a taxa SELIC se manteve em patamar elevado e comparativamente estável em relação aos demais regimes examinados neste capítulo, ao passo que o IPCA seguia trajetória de moderação. Tal configuração macroeconômica, juros altos, porém sem rupturas abruptas de política monetária, constitui o ambiente sobre o qual recai a avaliação quantitativa central deste trabalho, conduzida sobre as duzentas e uma execuções independentes do procedimento de treinamento descritas na seção 3.6.2, em virtude da qual este cenário recebe exame mais detalhado que os demais. As subseções seguintes examinam, em ordem, as estatísticas descritivas agregadas do desempenho do modelo (4.2.1), sua confrontação inferencial com as estratégias de referência (4.2.2), a distribuição dos resultados entre as execuções (4.2.3) e o comportamento operacional e a trajetória de capital do modelo (4.2.4), encerrando-se com uma síntese das evidências do cenário (4.2.5).

4.2.1 Estatísticas Descritivas Agregadas

A Tabela 5 consolida as estatísticas descritivas do modelo PPO e das três estratégias de referência no cenário principal.

Tabela 5 - Estatísticas descritivas do cenário principal (2023–2024)

Métrica	PPO (n = 201)	Buy & Hold	SELIC	Risk Parity
Retorno total (%)	25,59	22,32	25,12	14,71
Retorno anualizado (%)	12,17	10,71	11,98	7,18
Volatilidade anualizada (%)	7,69	6,66	0,08	4,64
Máximo drawdown (%)	–5,26	–3,62	0	–6,33
Índice de Sharpe	0,044	–0,147	–	–0,936
Índice de Sortino	0,071	–0,200	–	–1,178

Índice de Calmar	2,49	2,96	–	1,13
Rebalanceamentos (média)	372,7	–	–	–
Custo total (R\$)	200,68	–	–	–

Fonte: Elaboração própria com base na análise empírica (2026)

O retorno total do modelo, agregado sobre as duzentas e uma execuções independentes, exhibe média de 25,59% e mediana de 25,64%, valores cuja proximidade sugere uma distribuição aproximadamente simétrica em torno da tendência central. A dispersão entre execuções, medida pelo desvio-padrão de 5,42 pontos percentuais, delimita uma faixa de desempenho que se estende de 10,85% a 40,32%, sem que qualquer execução incorra em prejuízo. Em termos comparativos, o retorno médio do modelo supera o da estratégia Buy & Hold (22,32%) e, com ampla margem, o do Risk Parity (14,71%), aproximando-se do rendimento acumulado da SELIC capitalizada no período (25,12%), paridade cuja significância estatística será examinada na subseção seguinte.

No que concerne ao risco, o modelo apresenta volatilidade anualizada média de 7,69% e drawdown máximo médio de –5,26%, configurando um perfil de oscilação moderada. O índice de Sharpe médio, contudo, situa-se em 0,044, valor próximo de zero que demanda interpretação cuidadosa. Por mensurar o retorno excedente sobre o ativo livre de risco por unidade de risco assumido (CAMPBELL; LO; MACKINLAY, 1997), um Sharpe quase nulo indica que, conquanto o retorno absoluto do modelo seja expressivo, seu excedente sobre a taxa SELIC, referência de custo de oportunidade incontornável em um regime de juros elevados, é apenas marginal. Em síntese, o modelo assume risco direcional de mercado para auferir rendimento estatisticamente equivalente ao do ativo livre de risco. Quanto à eficiência risco-retorno aferida pelo índice de calmar, o modelo registra média de 2,49, inferior à do Buy & Hold (2,96), o que antecipa uma vulnerabilidade estrutural, a assunção proporcionalmente maior de risco de cauda, discutida em profundidade na seção 4.6.

No plano operacional, cada execução do modelo realiza, em média, 372,7 rebalanceamentos ao longo do período de teste, o que caracteriza uma política transaccionalmente ativa. O custo cumulativo médio associado a essa atividade é de R\$ 200,68, equivalente a aproximadamente dois por cento do capital inicial de R\$ 10.000 ao longo do biênio avaliado, ou cerca de um por cento ao ano. Todos os

indicadores de retorno reportados nesta e nas demais subseções correspondem a valores líquidos, calculados após a dedução dos custos transacionais incorridos por cada execução, de modo que o desempenho do modelo não se beneficia de qualquer abstração da fricção operacional.

4.2.2 Análise Inferencial vs. Baselines

A confrontação do desempenho do modelo com as estratégias de referência foi conduzida por meio de testes t de uma amostra. A escolha decorre da natureza assimétrica das quantidades comparadas: ao passo que o modelo PPO produz uma distribuição de retornos, uma observação por execução, num total de duzentas e uma, cada estratégia de referência é determinística sobre o conjunto de teste, resultando em um único valor. O teste t de uma amostra confronta, portanto, a média da distribuição de retornos do modelo contra cada valor de referência fixo, sob a hipótese nula de igualdade. A Tabela 6 reporta, para cada par, a estatística t, o p-valor e o intervalo de confiança de 95% para a diferença de médias.

Tabela 6 – Análise inferencial do desempenho frente às estratégias de referência (cenário principal, 2023 - 2024)

Comparação	Diferença média (p.p.)	Estatística t	p-valor	IC 95% da diferença (p.p.)	Superação (%)
PPO vs Buy & Hold	3,27	8,54	$\approx 3 \times 10^{-15}$	[2,51; 4,02]	69,7
PPO vs SELIC	0,47	1,22	0,22	[-0,29; 1,22]	54,2
PPO vs Risk Parity	10,88	28,42	$\approx 4 \times 10^{-72}$	[10,12; 11,63]	98,5

Fonte: *Elaboração própria (2026)*

Cumpramos delimitar o alcance dessa inferência: como a variabilidade entre as duzentas e uma observações decorre exclusivamente da inicialização aleatória do procedimento de treinamento (HENDERSON et al., 2018), e não de múltiplas realizações de mercado, os testes aferem se o resultado esperado do procedimento difere das referências sobre um único caminho histórico out-of-sample, não constituindo medida da capacidade de generalização do modelo a outras trajetórias de mercado.

Os resultados indicam superação do Buy & Hold com significância estatística robusta: a diferença média de 3,27 pontos percentuais corresponde a uma estatística t de +8,54 ($p \approx 3 \times 10^{-15}$), com intervalo de confiança de 95% para a diferença situado

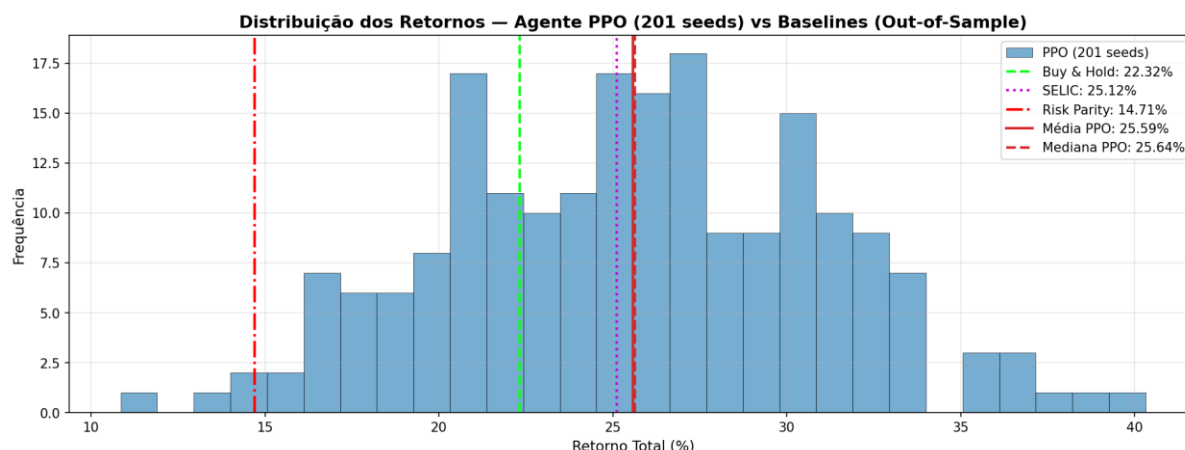
entre 2,51 e 4,02 pontos percentuais. A confrontação com o Risk Parity revela dominância ainda mais expressiva ($t = +28,42$; $p \approx 4 \times 10^{-72}$). Já o cotejo com a SELIC não permite rejeitar a hipótese de igualdade (diferença média de 0,47 ponto percentual; $t = +1,22$; $p = 0,22$), configurando empate técnico, resultado coerente com o índice de Sharpe quase nulo discutido na subseção anterior, porquanto o modelo, ao igualar o rendimento do ativo livre de risco, não gera excedente estatisticamente distinguível. As taxas de superação por execução corroboram o quadro: o modelo supera o Buy & Hold em 69,7% das execuções, a SELIC em 54,2% e o Risk Parity em 98,5%, sendo o retorno positivo em 100% delas. Registre-se, por fim, uma ressalva: a magnitude da significância na comparação com o Buy & Hold é amplificada pelo tamanho amostral, duzentas e uma execuções reduzem substancialmente o erro-padrão da média, de modo que o p-valor diminuto atesta a consistência do resultado entre inicializações, e não a dimensão econômica da vantagem, adequadamente representada pela diferença de 3,27 pontos percentuais.

A análise inferencial estende-se à eficiência risco-retorno, dimensão em que o quadro se inverte. Conforme antecipado, o modelo não supera o Buy & Hold no índice de Calmar: a média de 2,49 ante os 2,96 da estratégia passiva traduz-se em diferença estatisticamente significativa em desfavor do modelo ($t = -7,11$; $p \approx 2 \times 10^{-11}$), e apenas 30,8% das execuções superam o Buy & Hold nessa métrica. O resultado evidencia que o retorno superior documentado nas comparações anteriores é obtido à custa de exposição proporcionalmente maior ao risco, o modelo entrega mais retorno por unidade de tempo, porém com drawdowns mais severos relativamente ao ganho. Trata-se da primeira manifestação de uma característica estrutural do método, cuja origem na função de recompensa e cujas implicações são examinadas na seção 4.6.

4.2.3 Distribuição dos Desempenhos

A Figura 5 apresenta o histograma dos retornos totais das duzentas e uma execuções, sobre o qual estão demarcados os valores das estratégias de referência.

Figura 5 - Histograma dos retornos das execuções - cenário principal (2023–2024)



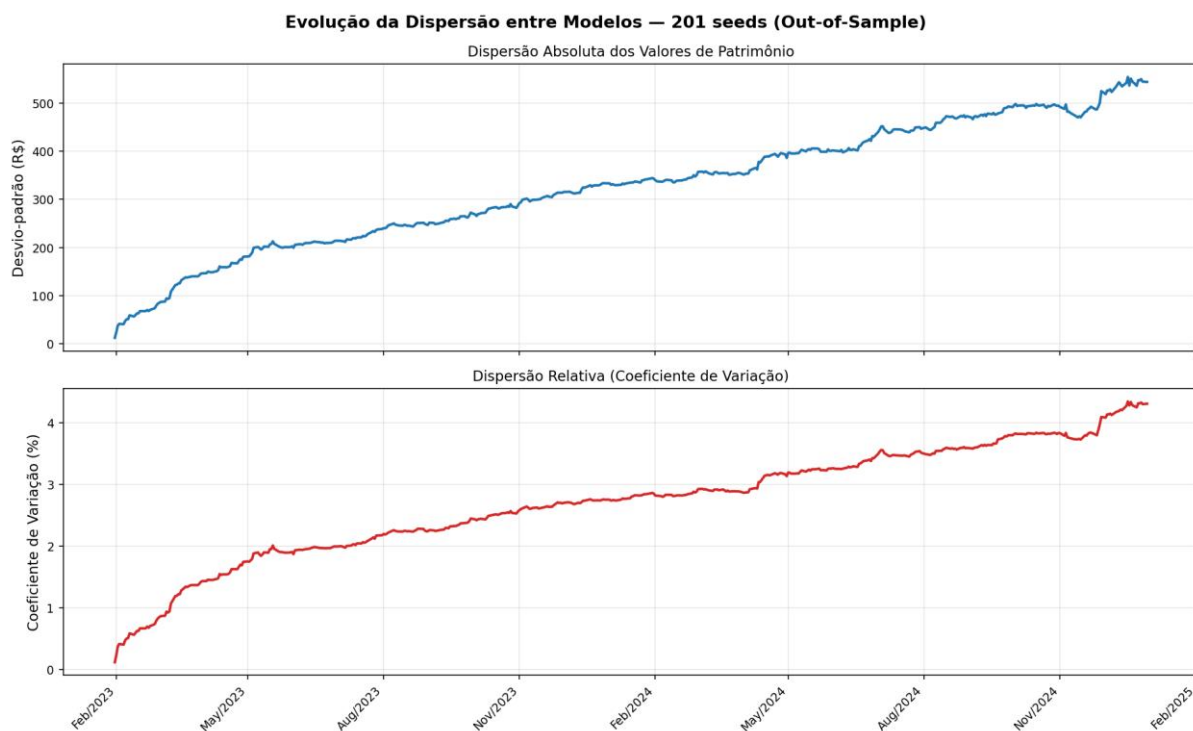
Fonte: Elaboração própria

A distribuição concentra-se predominantemente entre 20% e 33%, com caudas pouco povoadas: à esquerda, afina-se até o mínimo de 10,85%; à direita, estende-se até o máximo de 40,32%. A coincidência entre média (25,59%) e mediana (25,64%) confirma a aproximada simetria já antevista nas estatísticas descritivas. Releva notar a ausência completa de massa em zona de prejuízo, nenhuma execução encerra o período com perda. O posicionamento dos baselines no eixo é igualmente informativo: o Buy & Hold (22,32%) localiza-se à esquerda do corpo principal da distribuição, a SELIC (25,12%) atravessa-o nas proximidades da mediana, e o Risk Parity (14,71%) situa-se na cauda esquerda extrema, abaixo de quase a totalidade das execuções.

O exame da distribuição revela, ademais, a virtual ausência de execuções degeneradas neste cenário. Apenas duas das duzentas e uma execuções (1,0%) realizam menos de cem rebalanceamentos no período, limiar adiante adotado para distinguir políticas ativas de degeneradas, e uma única encontra-se efetivamente congelada. Ainda assim, mesmo essas execuções permanecem rentáveis, com retorno médio de 19,60%, ante 25,65% das demais. A raridade do fenômeno contrasta de maneira acentuada com os cenários de estresse examinados adiante, em especial o de aperto monetário, no qual a degeneração acomete um terço das execuções; o contraste sugere que o regime de juros estáveis constitui ambiente propício à convergência do procedimento de treinamento para políticas ativas e funcionais.

A Figura 6 documenta a evolução da dispersão entre as execuções ao longo do período de teste, em dois painéis.

Figura 6 - Evolução da dispersão entre execuções - cenário principal (2023–2024)



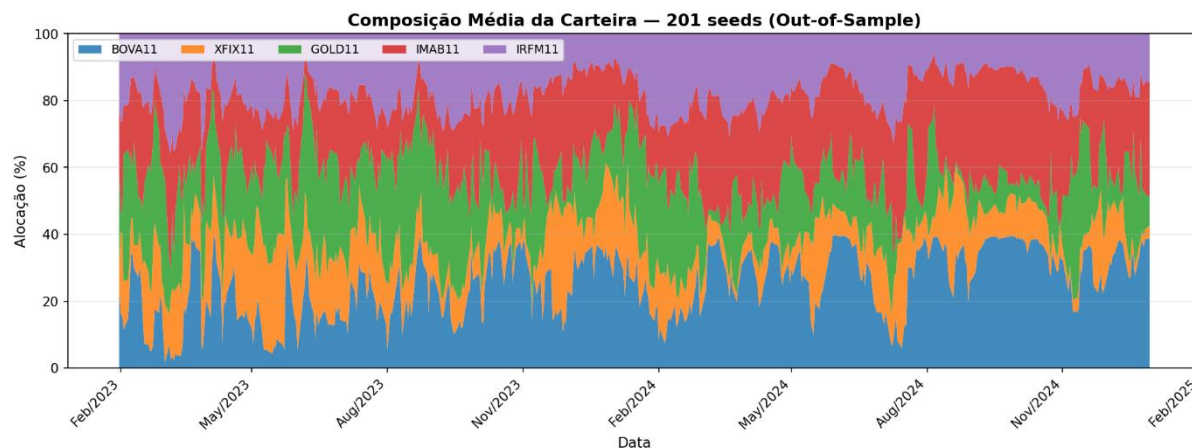
Fonte: Elaboração própria (2026)

O painel superior, referente ao desvio-padrão absoluto dos valores de patrimônio, exhibe crescimento monotônico, partindo de aproximadamente R\$ 10 no início do período e alcançando cerca de R\$ 540 ao final. Tal crescimento, isoladamente, poderia ser atribuído a mero efeito de escala, dado que carteiras de maior valor tendem a apresentar maior dispersão absoluta. É o painel inferior, contudo, que dirime a ambiguidade: o coeficiente de variação, medida de dispersão relativa, imune ao efeito de escala, também cresce de forma monotônica, de cerca de 0,1% para 4,3% ao longo do período. A elevação da dispersão relativa indica que as execuções não apenas se distanciam em termos absolutos, mas divergem genuinamente em trajetória, refletindo a sensibilidade do desempenho à inicialização aleatória do treinamento (HENDERSON et al., 2018). O achado fornece, retrospectivamente, a justificativa para o protocolo multi-seed, evidenciando que uma avaliação fundada em execução única estaria sujeita a localizar-se em qualquer ponto de uma banda de desempenho que se alarga progressivamente, o que comprometeria a representatividade da inferência.

4.2.4 Trajetória e Comportamento Operacional

A Figura 7 exibe a composição média da carteira do modelo ao longo do período, agregada sobre as duzentas e uma execuções.

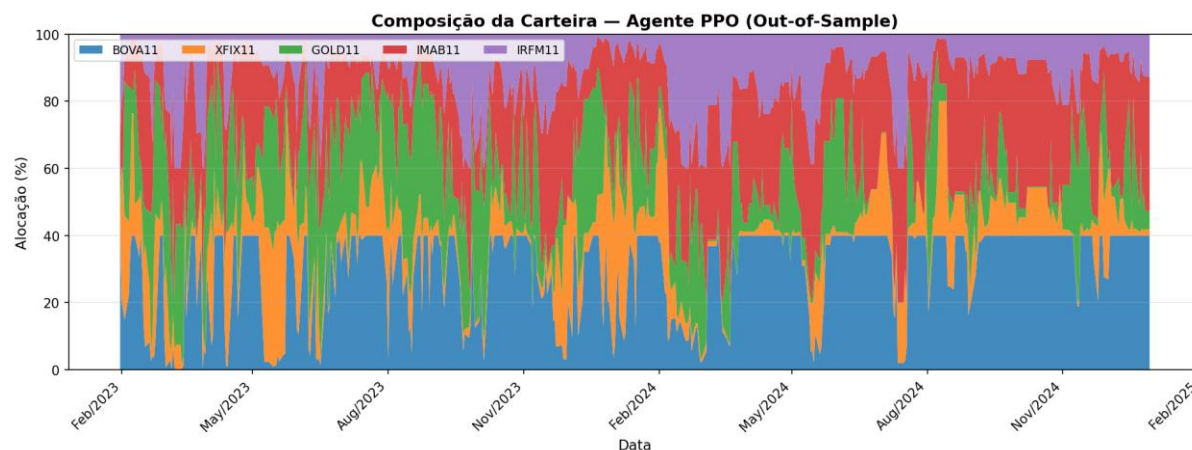
Figura 7 - Composição média da carteira do modelo - cenário principal (2023–2024)



Fonte: Elaboração própria (2026)

A alocação média revela-se acentuadamente diversificada e próxima da equiponderação, com os cinco ativos repartindo-se em faixas que vão de 13,9% (XFIX11) a 25,2% (BOVA11), sem predominância marcante de qualquer classe. Essa aparência equilibrada, contudo, não corresponde à política executada por qualquer execução individual: como evidencia a Figura 8, que ilustra a composição de uma execução representativa.

Figura 8 - Composição representativa de uma carteira do modelo - cenário principal (2023 - 2024)

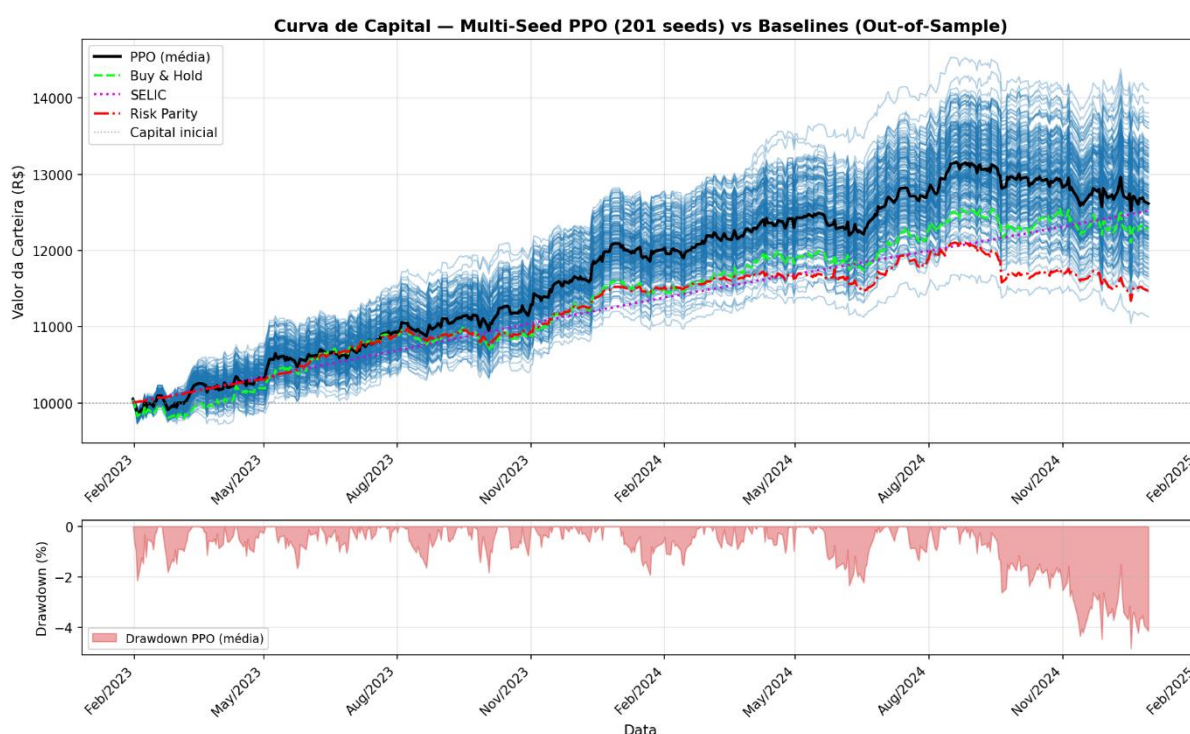


Fonte: Elaboração própria (2026)

Cada modelo opera de modo intensamente ativo e concentrado, deslocando-se com frequência diária entre os ativos e amiúde saturando o limite de concentração de 40%. A suavidade da composição média é, portanto, artefato da agregação de políticas heterogêneas, coerente com a divergência entre execuções documentadas na seção 4.2.3, e não reflexo de uma estratégia passiva comum às sementes. A figura individual é apresentada a título ilustrativo do comportamento operacional típico, não como objeto de inferência estatística.

A Figura 9 apresenta a curva de capital das duzentas e uma execuções, com a trajetória média destacada e confrontada às estratégias de referência.

Figura 9 - Curva de capital multi-seed e estratégias de referência - cenário principal (2023–2024)



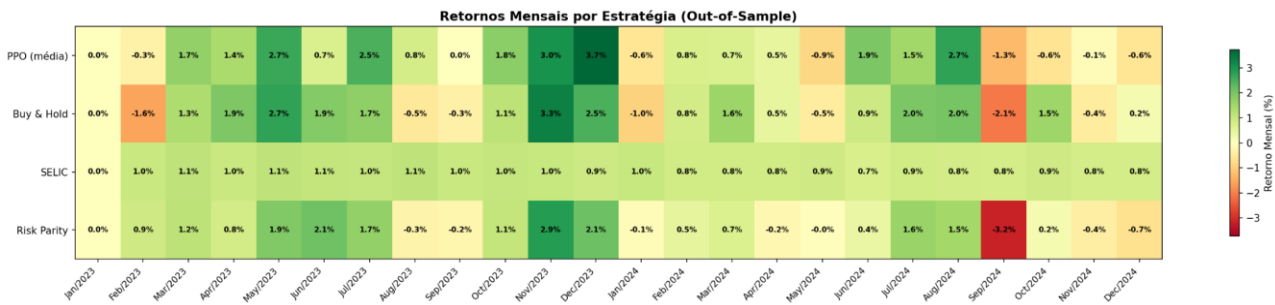
Fonte: Elaboração própria (2026)

A trajetória média do modelo posiciona-se acima das curvas do Buy & Hold e da SELIC durante a maior parte do período, distanciamento que se acentua a partir do segundo semestre de 2023 e atinge seu ápice em meados de 2024. Cumpre, todavia, registrar que essa vantagem se comprime no trecho final: ao longo do último trimestre de 2024, a margem do modelo sobre as estratégias passivas estreita-se sensivelmente, refletindo um período de desempenho relativo desfavorável que será

retomado na análise de drawdown. A banda formada pelo conjunto das execuções, por sua vez, alarga-se progressivamente ao redor da média, representação visual direta da dispersão crescente analisada na subseção anterior.

A Figura 10 dispõe, em mapa de calor, os retornos mensais de cada estratégia ao longo do biênio.

Figura 10 - Mapa de calor dos retornos mensais - cenário principal (2023–2024)

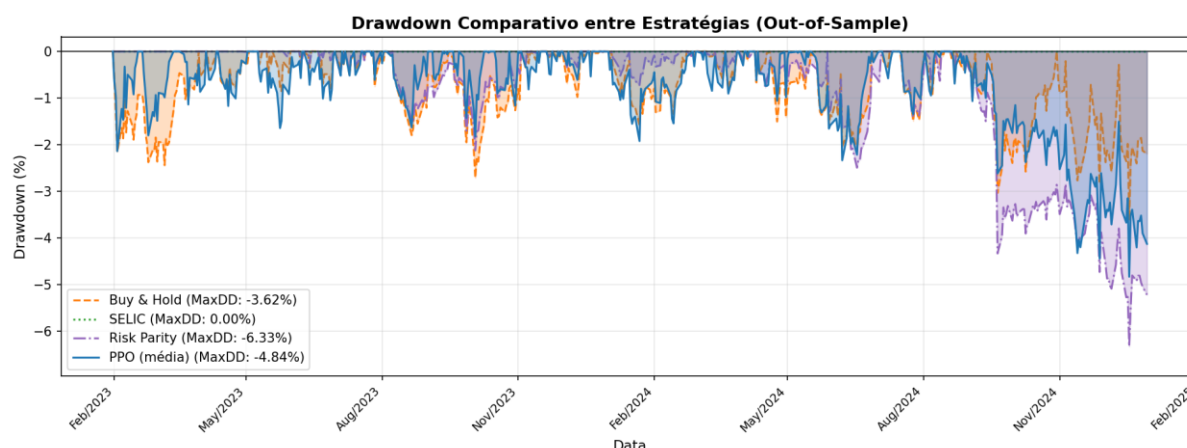


Fonte: Elaboração própria (2026)

A leitura mês a mês sustenta o argumento de que a vantagem cumulativa do modelo não se origina de poucos meses excepcionais, mas distribui-se ao longo do período. O modelo médio supera o Buy & Hold em cerca de metade dos meses avaliados, sem que nenhum mês isolado responda desproporcionalmente pela diferença acumulada. Parcela relevante dessa vantagem provém, ademais, da contenção de perdas em meses de baixa, nos quais o modelo recua menos que a estratégia passiva, padrão que, registre-se, convive com drawdowns diários mais profundos, conforme se examina a seguir, uma vez que a agregação mensal atenua as quedas entre os meses capturadas na frequência diária.

Por fim, a Figura 11 confronta as curvas de drawdown das estratégias.

Figura 11- Curvas de drawdown das estratégias - cenário principal (2023–2024)



Fonte: Elaboração própria (2026)

O modelo exibe drawdown máximo médio de $-5,26\%$ por execução, ao passo que a curva de patrimônio média registra $-4,84\%$, diferença que merece esclarecimento metodológico: como cada execução atinge seu vale em momento distinto, a média entre trajetórias atenua os recuos, de modo que o drawdown da série média subestima o risco efetivamente experimentado por uma execução típica. Em qualquer das medidas, o modelo incorre em drawdowns mais profundos e frequentes que o Buy & Hold ($-3,62\%$), embora menos severos que os do Risk Parity ($-6,33\%$). A vulnerabilidade concentra-se notadamente no último trimestre de 2024, quando o modelo experimenta seu recuo mais acentuado do período, observação que converge com a compressão de desempenho identificada na curva de capital e reitera a exposição ao risco de cauda já assinalada na análise do índice de calmar.

4.2.5 Síntese do Cenário Principal

Em síntese, o cenário principal evidencia o desempenho mais favorável do modelo entre os regimes avaliados. Sob condições de juros elevados e estáveis, o modelo PPO superou a estratégia passiva de Buy & Hold com significância estatística robusta, dominou amplamente o Risk Parity e equiparou-se ao rendimento da SELIC, mantendo retorno positivo na totalidade das execuções. Tal superioridade, contudo, não se reproduz nas dimensões de risco ajustado. O índice de Sharpe quase nulo revela que o excedente do modelo sobre o ativo livre de risco é apenas marginal, e o índice de calmar inferior ao do Buy & Hold, superado por menos de um terço das

execuções, denuncia que o retorno adicional é obtido mediante exposição proporcionalmente maior ao risco. A esses limites soma-se a dispersão crescente entre execuções, que impõe substancial incerteza sobre o desempenho de uma realização individual ao longo do tempo. O quadro que emerge é, portanto, o de um modelo competitivo, porém não dominante em todas as métricas, ainda sob o regime que lhe é mais propício. Resta averiguar se essa competitividade se sustenta em contextos macroeconômicos adversos, questão a que se dedicam as duas seções seguintes, que submete o modelo a um choque agudo de mercado e a um ciclo de aperto monetário.

4.3 CENÁRIO DE ESTRESSE PANDÊMICO: COVID-19 (2020)

O primeiro cenário de estresse submete o modelo ao exercício de 2020, marcado pelo choque pandêmico da COVID-19. O regime macroeconômico do período distingue-se radicalmente do cenário principal. A taxa SELIC foi reduzida a sua mínima histórica, ao passo que o mercado acionário sofreu queda abrupta no primeiro trimestre, seguida de recuperação ao longo dos meses subsequentes. A avaliação neste e no próximo cenário, em conformidade com o protocolo da seção 3.7, foi conduzida sobre amostra reduzida de trinta execuções independentes, destinada à aferição de robustez sob condições adversas. Em consequência, os testes inferenciais aqui reportados devem ser interpretados como evidência indicativa, e não conclusiva, sendo lidos em conjunto com a análise descritiva da distribuição dos desempenhos.

A Tabela 7 consolida as estatísticas do cenário.

Tabela 7 - Estatísticas descritivas do cenário de estresse pandêmico (COVID-19, 2020)

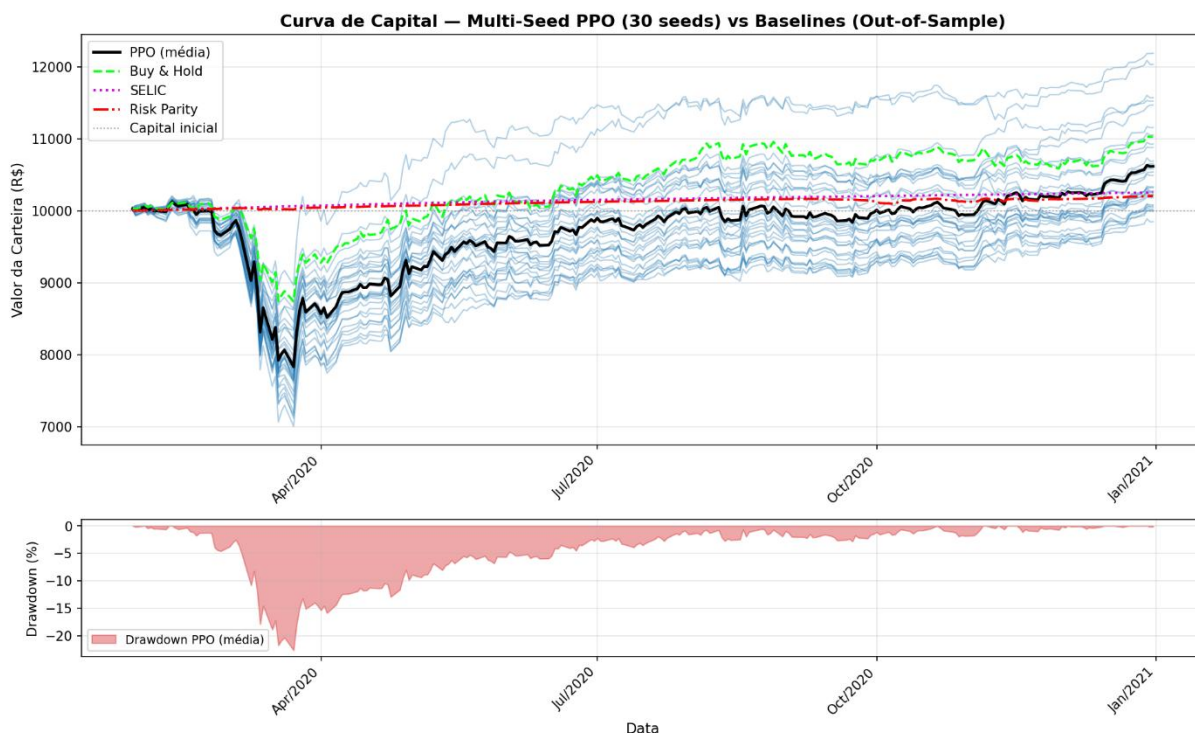
Métrica	PPO (n = 30)	Buy & Hold	SELIC	Risk Parity
Retorno total (%)	5,9	9,92	2,51	2,08
Retorno anualizado (%)	6,2	10,44	2,64	2,18
Volatilidade anualizada (%)	20,12	14,73	0,05	0,83
Máximo drawdown (%)	-22,8	-13,62	0	-0,66
Índice de Sharpe	0,26	0,56	—	-0,69
Índice de Sortino	0,36	0,74	—	-0,98
Índice de Calmar	0,33	0,77	—	3,3

Rebalanceamentos (média)	158,7	–	–	–
Custo total (R\$)	78,93	–	–	–

(Fonte: Elaboração própria com base na análise empírica)

O modelo registra retorno médio de 5,90% e mediana de 3,92%, ficando atrás do Buy & Hold (9,92%) por margem estatisticamente significativa ($t = -3,53$; $p = 0,0014$; intervalo de confiança de 95% para a diferença entre $-6,36$ e $-1,69$ pontos percentuais). A inferioridade reproduz-se em todas as dimensões de risco: o índice de Sharpe médio (0,27) corresponde a aproximadamente metade do registrado pela estratégia passiva (0,56); o drawdown máximo médio ($-22,83\%$) supera com folga o do Buy & Hold ($-13,62\%$); e o índice de Calmar (0,33) situa-se igualmente abaixo (0,77). Ainda assim, o modelo preserva ampla superioridade sobre as alternativas de renda fixa, superando tanto a SELIC (2,51%; $t = +2,97$; $p = 0,006$) quanto o Risk Parity (2,08%; $t = +3,34$; $p = 0,002$), e encerra o período com retorno positivo em 86,7% das execuções. O comportamento dos ativos esclarece o resultado: o ouro disparou ($+54,3\%$), enquanto a renda variável ($-6,2\%$) e os fundos imobiliários ($-20,8\%$) sofreram perdas pronunciadas. A carteira diversificada do modelo foi, assim, simultaneamente penalizada pela exposição aos ativos de risco e amparada pela posição em ouro, do que resulta desempenho positivo, porém inferior ao da estratégia passiva equiponderada, que capturou de forma mais direta a recuperação do conjunto do mercado. A curva de capital multi-seed (Figura 12) ilustra esse descompasso. A trajetória média do modelo corre abaixo da do Buy & Hold ao longo de quase todo o período de recuperação, mantendo-se, contudo, acima das estratégias de renda fixa.

Figura 12 - Curva de capital multi-seed e estratégias de referência - cenário COVID-19 (2020)

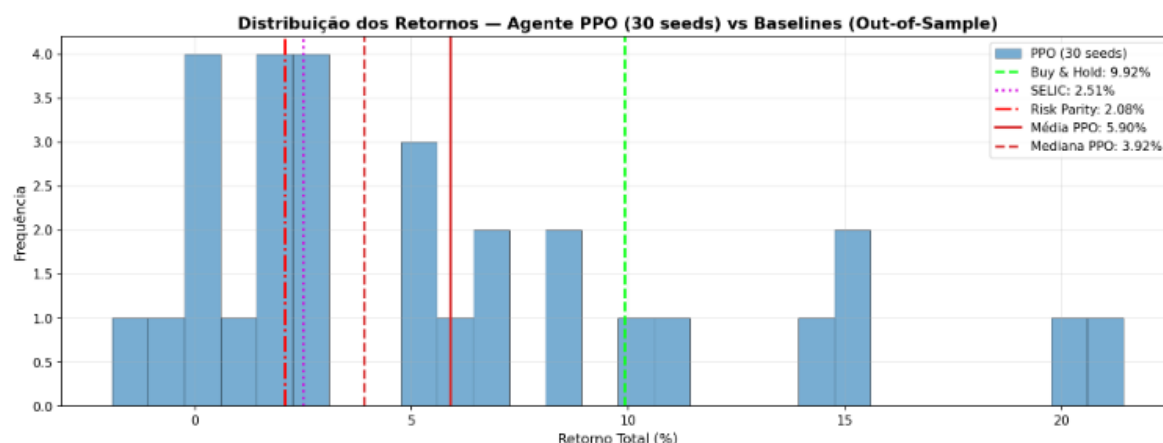


Fonte: Elaboração própria (2026)

A análise da distribuição revela um padrão de degeneração qualitativamente distinto do que se observará no cenário de aperto monetário. Sete das trinta execuções (23,3%) realizaram menos de cem rebalanceamentos, proporção bem superior à do cenário principal. Tal degeneração, contudo, mostra-se benigna sob a ótica do desempenho. As execuções degeneradas obtiveram retorno médio de 5,03%, valor estatisticamente indistinguível do das execuções ativas (6,16%), e a correlação entre o número de rebalanceamentos e o retorno é nula ($r = 0,094$; $p = 0,62$). A explicação reside na natureza do regime: em uma recuperação em "V", a inação, manter a composição sem rebalancear, não penaliza o resultado, pois permite acompanhar o repique do mercado. O histograma dos retornos (Figura 13) exhibe a distribuição das execuções concentrada em valores positivos, sem a separação em dois grupos que caracterizará o cenário de aperto monetário, leitura coerente com o caráter benigno da degeneração aqui observada. As demais visualizações do cenário, o mapa de calor mensal e a evolução da dispersão entre execuções, de natureza complementar, constam do Apêndice A. O padrão contrasta diretamente com o do

regime de aperto monetário, analisado a seguir, em que a mesma degeneração assume caráter severamente prejudicial.

Figura 13 - Histograma dos retornos das execuções - cenário COVID-19 (2020)



Fonte: Elaboração própria (2026)

O cenário pandêmico fornece, em síntese, a evidência mais nítida do perfil comportamental do modelo. Submetido a um choque agudo, o modelo não atuou como gestor defensivo. O drawdown máximo médio de -22,83%, mais profundo que o da própria estratégia passiva, e não logrou capturar a recuperação subsequente com a mesma eficiência do Buy & Hold. Por outro lado, tampouco se refugiou na renda fixa, manteve exposição aos ativos de risco e, com isso, superou amplamente as alternativas conservadoras e preservou retorno positivo na maioria das execuções. Configura-se, assim, um alocador de perfil direcional, que assume risco de mercado de forma persistente e dele extrai retorno mesmo em condições adversas, mas que carece de mecanismos de proteção patrimonial em episódios de queda abrupta, caracterização retomada e aprofundada na seção 4.6.

4.4 CENÁRIO DE EXPANSÃO DA SELIC (2021 - 2022)

O segundo cenário de estresse compreende os anos de 2021 e 2022, período do mais intenso ciclo de aperto monetário do histórico recente brasileiro: a taxa SELIC elevou-se de 2% a 13,75% no intervalo, enquanto o IPCA atingiu pico de aproximadamente 12% antes de iniciar reversão. Trata-se de um regime de transição macroeconômica abrupta, distinto tanto da estabilidade do cenário principal quanto do

choque agudo e pontual da pandemia. A avaliação manteve a amostra de trinta execuções independentes, nos termos da ressalva metodológica enunciada na seção anterior.

A Tabela 8 sintetiza os resultados.

Tabela 8 - Estatísticas descritivas do cenário de aperto monetário (expansão da SELIC, 2021-2022)

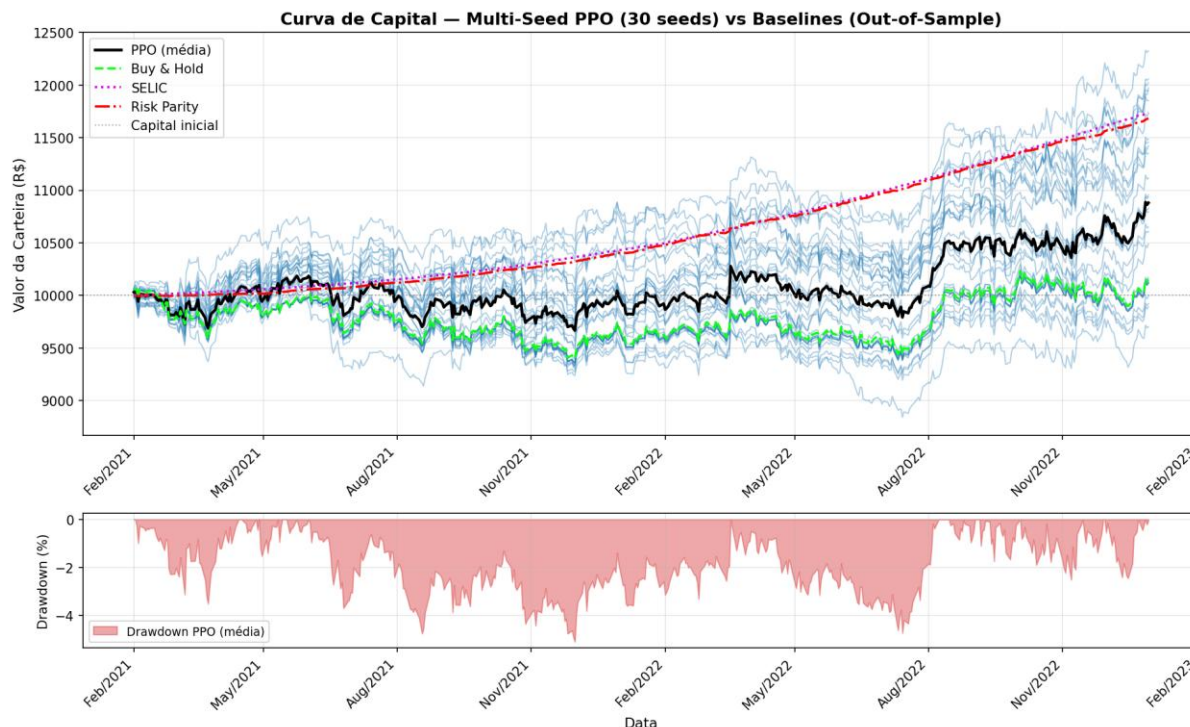
Métrica	PPO (n = 30)	Buy & Hold	SELIC	Risk Parity
Retorno total (%)	8,48	1	17,39	16,89
Retorno anualizado (%)	4,13	0,5	8,43	8,2
Volatilidade anualizada (%)	8,96	7,55	0,25	0,45
Máximo drawdown (%)	-7,41	-6,55	0	-0,06
Índice de Sharpe	-0,416	-0,934	—	0,108
Índice de Sortino	-0,565	-1,256	—	0,171
Índice de Calmar	0,62	0,08	—	—
Rebalanceamentos (média)	282,6	—	—	—
Custo total (R\$)	148,15	—	—	—

Fonte: Elaboração própria (2026)

O modelo registra retorno médio de 8,48% e mediana de 7,93%, com elevada dispersão entre execuções (desvio-padrão de 8,10 pontos percentuais) e, fato que distingue este cenário dos demais, índice de Sharpe médio negativo (-0,416). O Sharpe negativo decorre diretamente do regime. Com a taxa SELIC em patamar muito elevado, o rendimento do modelo situou-se abaixo do ativo livre de risco, tornando negativo o excedente que o índice mensura. A confrontação inferencial revela um quadro nítido de robustez condicional: o modelo supera o Buy & Hold, que praticamente estagnou no período (1,00%; $t = +5,06$; $p \approx 2 \times 10^{-5}$), porém é dominado de modo significativo tanto pela SELIC (17,39%; $t = -6,02$; $p \approx 1 \times 10^{-6}$) quanto pelo Risk Parity (16,89%; $t = -5,69$; $p \approx 4 \times 10^{-6}$). O desempenho dos ativos esclarece cabalmente o resultado: apenas a renda fixa foi rentável no período (IRFM11 +18,0%; IMAB11 +4,7%), ao passo que a renda variável (-12,6%), os fundos imobiliários (-11,0%) e o ouro (-7,9%) acumularam perdas. As estratégias vencedoras, a SELIC, por construção, e o Risk Parity, que pondera os ativos inversamente à volatilidade e por isso se concentra na renda fixa de baixa oscilação, beneficiaram-se exatamente dos únicos ativos rentáveis, ao passo que a alocação diversificada do modelo preservou exposição substancial às classes que se desvalorizaram. A curva de capital multi-seed (Figura 14) evidencia esse descompasso: a trajetória média do modelo,

ainda que superior à do Buy & Hold, permanece ao longo de todo o período de teste abaixo das estratégias de renda fixa, que capitalizaram diretamente a elevação dos juros.

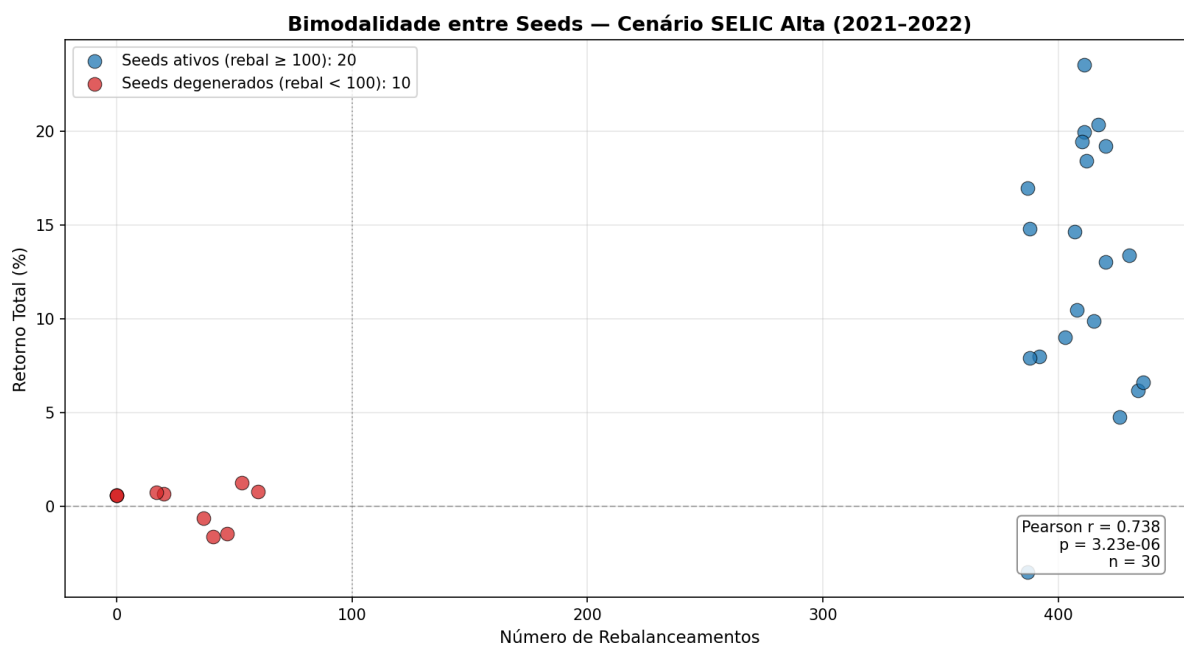
Figura 14 - Curva de capital multi-seed e estratégias de referência - cenário de expansão da SELIC (2021–2022)



Fonte: Elaboração própria (2026)

Dois aspectos do cenário merecem exame detido. O primeiro é a inversão da relação de eficiência risco-retorno observada no cenário principal. Aqui, o modelo supera o Buy & Hold no índice de Calmar (0,62 contra 0,08; $t = +4,75$; $p \approx 5 \times 10^{-5}$), sendo-lhe superior em 66,7% das execuções. A aparente contradição com o cenário principal, no qual o modelo era inferior nessa métrica, dissolve-se ao se observar que, no aperto monetário, a estratégia passiva não apenas estagnou em retorno como o fez incorrendo em drawdown, de modo que a vantagem do modelo em Calmar reflete menos um mérito próprio do que o fracasso do referente passivo. O segundo aspecto é a bimodalidade severa da distribuição, ilustrada na Figura 15, que dispõe o retorno de cada execução contra o respectivo número de rebalanceamentos.

Figura 15 - Retorno por execução versus número de rebalanceamentos (bimodalidade) - cenário de expansão da SELIC (2021-2022)



Fonte: Elaboração própria (2026)

Distinguem-se dois agrupamentos: 67% das execuções permanecem ativas e obtiveram um retorno médio de 12,65%, ao passo que 33% degeneram (menos de cem rebalanceamentos) e rendem em média apenas 0,15%, com três execuções inteiramente congeladas. É nesse segundo agrupamento que se concentram as poucas execuções de retorno negativo, de sorte que o retorno permanece positivo em 86,7% do total. A correlação entre rebalanceamentos e retorno é forte e altamente significativa ($r = 0,738$; $p \approx 3 \times 10^{-6}$). Ressalte-se que essa bimodalidade acentuada atenua a adequação da suposição de normalidade subjacente ao teste t, de sorte que as conclusões inferenciais devem ser ponderadas à luz desta análise de distribuição, da qual extraem sua sustentação principal. Esse comportamento é o oposto da degeneração benigna observada na pandemia. Lá, a falta de rebalanceamento não prejudicava o resultado, pois a recuperação em "V" recompensava a simples permanência no mercado. Aqui, ao contrário, como a alta dos juros exigia migração ativa para a renda fixa, essa mesma inação condena a execução à estagnação. As demais visualizações do cenário, o histograma dos retornos, o mapa de calor mensal e a evolução da dispersão entre execuções, de natureza complementar, constam do Apêndice A e corroboram o diagnóstico.

O regime de aperto monetário expõe a fragilidade do método sob transição macroeconômica abrupta: a elevada incidência de degeneração compromete a confiabilidade de uma execução individual e, mantendo alocação diversificada, o modelo não migra para a renda fixa que o regime recompensava, limitação cuja origem é examinada na seção 4.6.

Em síntese, o cenário de aperto monetário delimita o polo inferior do desempenho condicional do modelo. Ele supera o Buy & Hold, que fracassou no período, mas é claramente dominado pelas estratégias que se beneficiaram diretamente da elevação dos juros, e expõe, na bimodalidade de seus resultados, a sensibilidade do procedimento de treinamento a regimes de transição abrupta. Conjugado aos cenários anteriores, o resultado sustenta a tese de que a robustez do modelo é condicional ao regime macroeconômico: competitivo sob juros estáveis, resiliente, porém inferior sob choque agudo, e frágil sob aperto monetário.

4.5 ANÁLISE COMPARATIVA ENTRE REGIMES

As seções precedentes examinaram o desempenho do modelo isoladamente em cada regime. A presente seção consolida essas evidências em análise comparativa, com o objetivo de averiguar a hipótese central deste trabalho, a robustez do desempenho do modelo diante de contextos macroeconômicos distintos, e de delimitar com precisão os limites dessa robustez.

A Tabela 9 reúne, em formato comparativo, os principais indicadores do modelo e das estratégias de referência nos três regimes.

Tabela 9 - Síntese comparativa do desempenho entre regimes macroeconômicos

Métrica	Principal (n = 201)	COVID (n = 30)	SELIC alta (n = 30)
Retorno PPO (%)	25,59	5,9	8,48
Retorno Buy & Hold (%)	22,32	9,92	1
Retorno SELIC (%)	25,12	2,51	17,39
Retorno Risk Parity (%)	14,71	2,08	16,89
Índice de Sharpe (PPO)	0,044	0,27	-0,416
Índice de Calmar (PPO)	2,49	0,33	0,62
Índice de Calmar (Buy & Hold)	2,96	0,77	0,08
Execuções positivas (%)	100	86,7	86,7
Degeneração (% das execuções)	1	23,3	33,3

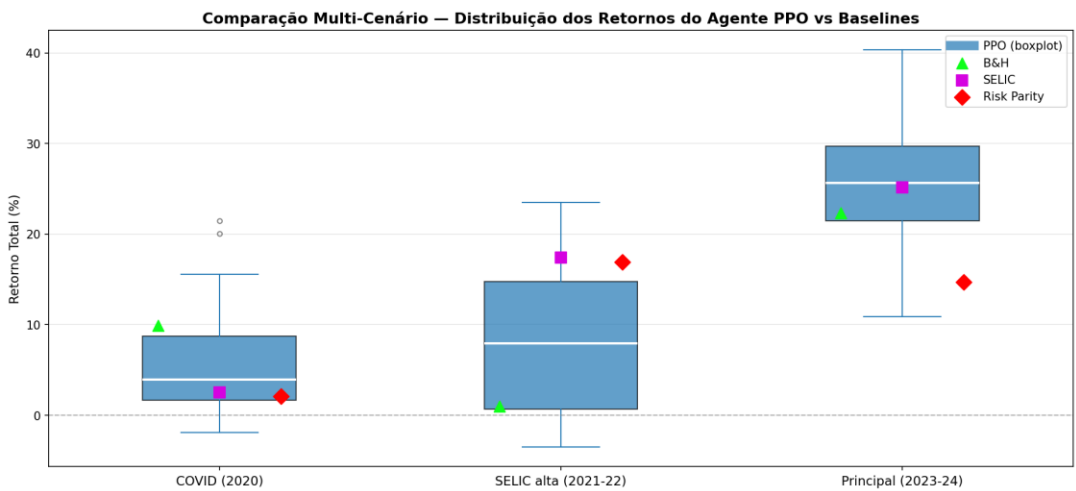
Veredito do PPO	supera B&H; $\approx$ SELIC; domina RP	perde p/ B&H; supera renda fixa	supera B&H; dominado por SELIC/RP
-----------------	---	------------------------------------	---

Fonte: Elaboração própria (2026)

O contraste evidencia um padrão de desempenho fortemente dependente do contexto. No regime de juros estáveis, o modelo alcança seu melhor desempenho relativo, superando o Buy & Hold, equiparando-se à SELIC e dominando o Risk Parity. No choque pandêmico, posiciona-se de forma intermediária: cede ao Buy & Hold, que captura mais eficientemente a recuperação do mercado, mas mantém ampla superioridade sobre as alternativas de renda fixa. No aperto monetário, ocupa a posição relativa mais desfavorável: ainda supera o Buy & Hold estagnado, porém é dominado pela SELIC e pelo Risk Parity, que se beneficiaram diretamente da elevação dos juros. Observa-se, ademais, que a identidade do melhor referente varia conforme o regime. A SELIC e o Risk Parity, dominados no cenário principal, tornam-se as estratégias superiores sob aperto monetário, o que reforça a natureza condicional de qualquer juízo de desempenho relativo.

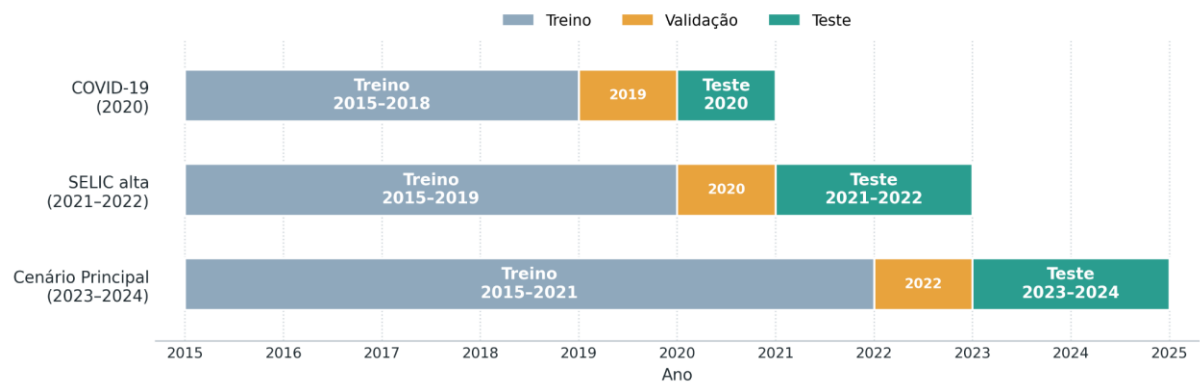
A Figura 16 dispõe, em diagramas de caixa lado a lado, a distribuição dos retornos do modelo nos três regimes, com os valores das estratégias de referência sobrepostas. Para situar temporalmente cada uma dessas distribuições, a Figura 17 apresenta a divisão dos dados em treino, validação e teste adotada em cada regime, evidenciando que o período de teste de cada cenário é integralmente posterior aos respectivos dados de treinamento e validação, em conformidade com o protocolo de avaliação out-of-sample descrito na seção 3.8.

Figura 16 - Distribuição dos retornos do modelo nos três regimes (diagramas de caixa)



Fonte: Elaboração própria (2026)

Figura 17 - Divisão temporal dos dados em treino, validação e teste por cenário



Fonte: Elaboração própria (2026)

A representação torna visualmente manifesta a assimetria de desempenho entre regimes: a caixa do cenário principal situa-se em patamar elevado e é comparativamente estreita, refletindo desempenho consistentemente superior. A do aperto monetário é a mais baixa e a mais dispersa, traduzindo graficamente a bimodalidade discutida na seção anterior. A do choque pandêmico ocupa posição intermediária. A posição relativa dos marcadores dos baselines em cada caixa recapitula, de forma sintética, as conclusões inferenciais das seções anteriores.

Dois padrões transversais merecem destaque por iluminarem a natureza condicional da robustez do modelo. O primeiro é a inversão do índice de Calmar entre regimes: no cenário principal, o modelo é inferior ao Buy & Hold nessa métrica (2,49

contra 2,96), ao passo que no aperto monetário lhe é superior (0,62 contra 0,08). A leitura conjunta dissolve a aparente incoerência, o modelo não é, em termos absolutos, nem melhor nem pior que a estratégia passiva quanto à eficiência risco-retorno. Ele é superior quando o referente passivo fracassa e inferior quando o referente prospera. O segundo padrão é o caráter condicional da degeneração de execuções: idêntico fenômeno estrutural, a convergência de parte das sementes para políticas inertes manifesta consequências opostas conforme o regime. No choque pandêmico a degeneração é benigna, sem correlação entre atividade e retorno ($r = 0,094$), pois a recuperação em "V" recompensa a mera permanência no mercado; no aperto monetário é maligna, com forte correlação ($r = 0,738$), pois o regime exigia realocação ativa que as sementes inertes não executaram. Ambos os padrões convergem para a mesma constatação: as propriedades do modelo não são intrínsecas, mas emergem da interação entre uma política fixa e as características de cada regime de mercado.

Resta confrontar diretamente a hipótese que motivou a incorporação das variáveis macroeconômicas ao espaço de estados: a de que o modelo adaptaria sua política de alocação ao regime vigente. Para tanto, computou-se a composição média da carteira, agregada sobre as sementes e o período, em cada cenário. A evidência é inequívoca quanto à invariância: a alocação média mantém-se próxima da equiponderação nos três regimes, com os pesos dos cinco ativos confinados a uma faixa estreita em torno de vinte por cento. A maior variação inter-regime não excede sete pontos percentuais, e três dos cinco ativos praticamente não se deslocam. O único movimento dotado de alguma coerência econômica é a redução gradual da exposição à renda variável (de 25,2% no regime estável a 20,1% no aperto), compatível com a contração de risco esperada sob elevação dos juros; tal deslocamento, porém, é modesto e vem acompanhado de um aumento da posição em fundos imobiliários no mesmo regime, movimento que contraria a sensibilidade negativa desses ativos aos juros e descaracteriza qualquer padrão adaptativo consistente. Para descartar que a uniformidade fosse artefato das sementes degeneradas, que, ao não rebalancear, preservam a composição equiponderada inicial, a análise foi repetida restringindo-se às sementes ativas; o padrão persiste com diferenças inferiores a um ponto percentual, confirmando que a invariância caracteriza também as políticas que efetivamente aprenderam a operar.

Conclui-se, portanto, que o modelo não adapta sua alocação ao regime macroeconômico. A robustez condicional documentada ao longo deste capítulo manifesta-se exclusivamente no plano do desempenho, e não no da política: o modelo executa, nos três regimes, uma alocação ativa e aproximadamente invariante, e a variação de seu desempenho decorre da interação dessa política estável com as condições de cada mercado, não de uma reconfiguração deliberada da carteira em resposta ao contexto. Cabe ressaltar que a composição média não corresponde à política de qualquer execução individual, por agregar políticas heterogêneas entre inicializações (HENDERSON et al., 2018), seu aspecto equilibrado oculta a atividade transacional concentrada de cada semente. As causas dessa ausência de adaptação e suas implicações são examinadas na seção seguinte.

4.6 DISCUSSÃO DOS RESULTADOS

Este capítulo encerra-se com a discussão integrada dos achados, que sintetiza interpretativamente as evidências dos três regimes e reconhece de modo explícito as limitações observadas no desempenho do modelo, bem como as direções de pesquisa que delas decorrem.

A hipótese que norteou o trabalho, a de que o aprendizado por reforço profundo constituiria abordagem viável e robusta para o rebalanceamento dinâmico de carteiras no mercado brasileiro, confirma-se de forma qualificada. O modelo PPO demonstrou viabilidade prática inequívoca: encerrou todos os regimes com retorno positivo na maioria das execuções e superou a estratégia passiva de Buy & Hold em dois dos três cenários avaliados, com significância estatística. A robustez, contudo, é condicional, e não absoluta. O modelo não domina uniformemente as alternativas, ele cede ao Buy & Hold sob o choque pandêmico e é claramente superado pelas estratégias de renda fixa sob o aperto monetário. A contribuição empírica do trabalho reside, portanto, menos em estabelecer a superioridade do método do que em delimitar com precisão as condições sob as quais ele é competitivo, e aquelas sob as quais não o é.

A primeira limitação, transversal aos cenários, concerne à eficiência risco-retorno. O modelo tende a entregar retorno mediante exposição proporcionalmente maior ao risco, como atesta o índice de Calmar inferior ao do Buy & Hold no cenário principal, superado por menos de um terço das execuções. Trata-se de característica

estrutural do método, e não de artefato: a função de recompensa adotada, o log-retorno líquido, premia o ganho sem penalizar explicitamente os drawdowns, de modo que o modelo não dispõe de incentivo para limitar perdas de pico a vale. A literatura de aprendizado por reforço aplicado a finanças oferece formulações alternativas de recompensa que incorporam o risco de forma explícita (ALMAHDI; YANG, 2017), cuja adoção constitui caminho natural para mitigar essa vulnerabilidade.

A segunda limitação manifesta-se na confiabilidade do procedimento de treinamento sob regimes adversos. No cenário de aperto monetário, um terço das execuções degenerou em políticas inertes, de retorno praticamente nulo, com forte correlação entre atividade transacional e retorno ($r = 0,738$). O fenômeno revela que, sob condições de mercado desfavoráveis, uma fração substancial das inicializações não converge para uma política funcional. Embora a documentação honesta dessa bimodalidade constitua, em si, achado científico legítimo sobre o comportamento do PPO nesta configuração, ela impõe cautela quanto à aplicação prática. A qualidade de uma execução individual não pode ser presumida, exigindo seleção ou validação criteriosa do modelo antes de qualquer emprego efetivo.

A terceira limitação, correlata, é a dispersão crescente entre execuções, documentada no cenário principal pela elevação monotônica do coeficiente de variação ao longo do tempo. A implicação prática é direta: para horizontes de aplicação longos, a incerteza sobre o desempenho de uma execução individual é substancial, pois o resultado de uma única semente pode situar-se em qualquer ponto de uma banda que se alarga progressivamente. Disso decorre que o protocolo de avaliação multi-seed adotado neste trabalho não constitui mero refinamento metodológico, mas necessidade prática para qualquer caracterização honesta do desempenho do método.

A invariância aproximada da política de alocação entre regimes, observada ainda que as variáveis macroeconômicas componham o espaço de estados, constitui resultado que requer interpretação cuidadosa, uma vez que parece contrariar a própria premissa que norteou o desenho do ambiente. A hipótese mais parcimoniosa para o fenômeno não está em uma incapacidade intrínseca do algoritmo, mas na heterogeneidade de escala das observações fornecidas à política. Como descrito na seção 3.4.1, os atributos que compõem o estado não foram submetidos a normalização e, por conseguinte, alimentam a rede em ordens de grandeza

marcadamente distintas: enquanto indicadores de risco-retorno, como o índice de Sortino, situam-se em escala de ordem unitária, a taxa SELIC diária ocupa a ordem de 10^{-4} . Em uma entrada não normalizada, tal disparidade faz com que a variação do sinal de juros seja numericamente ofuscada pelas demais variáveis, tornando inacessível à política, na prática, justamente a informação que deveria fundamentar a adaptação ao regime macroeconômico.

Decorre dessa disparidade de escala uma explicação mecanística coerente para o achado: privado de um sinal de juros efetivamente acessível, o modelo converge para uma política ativa, porém insensível ao regime. Sob essa leitura, a condicionalidade do desempenho documentada nas seções anteriores não decorre de adaptação deliberada, mas da interação entre uma política aproximadamente estável e as condições idiossincráticas de cada regime de mercado. Longe de invalidar os resultados, o diagnóstico delimita com precisão o escopo das conclusões e fundamenta a principal direção de trabalho futuro: a reavaliação da hipótese adaptativa sob um espaço de estados normalizado, procedimento cujo impacto sobre o desempenho de algoritmos de gradiente de política é amplamente documentado na literatura (HUANG et al., 2022).

Estabelecido que o desempenho superior do modelo no cenário principal não decorre de adaptação macroeconômica, resta identificar sua fonte. A análise da composição média revela que o modelo sobrepondera a renda variável (BOVA11, 25,2%) relativamente à distribuição equiponderada, mantendo posição neutra a levemente subponderada no ouro (GOLD11, 19,3%). Confrontados esses pesos com o desempenho dos ativos no período, em que o BOVA11 foi o ativo de menor retorno (aproximadamente 8%) e o GOLD11 o de maior, por larga margem (superior a 60%), constata-se que o tilt estratégico médio do modelo foi, na realidade, desfavorável: a carteira inclinou-se para o ativo de pior rentabilidade e afastou-se do de melhor. A superação observada sobre as estratégias de referência não pode, portanto, ser atribuída a um posicionamento estratégico vantajoso da composição. Sua origem reside necessariamente na realocação dinâmica de curto prazo executada ao longo do período, materializada no intenso comportamento transacional (média de 372,7 rebalanceamentos por execução). O modelo não atua, assim, como estrategista que define uma composição-alvo favorável, mas como operador tático que extrai retorno do reposicionamento de alta frequência entre os ativos, capacidade distinta da

adaptação macroeconômica entre regimes e dela independente. Registre-se, por fim, que essa vantagem tática, por depender de reposicionamento frequente, é a componente do desempenho mais sensível à estrutura de custos transacionais, e sua robustez sob hipóteses de custo mais conservadoras permanece como questão em aberto.

As limitações identificadas delineiam um programa de pesquisa subsequente. A direção de maior prioridade é a normalização das observações, que, ao tornar o sinal macroeconômico efetivamente acessível à política, permitiria reavaliar de forma justa a hipótese adaptativa central deste trabalho. Em segundo lugar, a substituição da função de recompensa por formulações que penalizam explicitamente o drawdown (ALMAHDI; YANG, 2017) poderia corrigir a vulnerabilidade do índice de Calmar. Recomenda-se, ademais, o desenvolvimento de mecanismos de detecção precoce de degeneração de sementes durante o treinamento; e o aprofundamento da decomposição entre alocação estratégica e timing tático em nível de execução individual, que esclareceria a consistência da vantagem tática identificada.

Em conclusão, o modelo PPO demonstra viabilidade prática condicional ao regime macroeconômico, com vantagens documentadas sob juros estáveis e limitações igualmente documentadas sob choque agudo e transição monetária. A caracterização que emerge do conjunto das evidências é precisa: o modelo é um alocador tático direcional, que extrai retorno do reposicionamento ativo de alta frequência, porém não é um gestor defensivo, pois não protege o patrimônio em quedas abruptas, nem um estrategista macroeconômico, pois não adapta sua alocação ao regime vigente. Essa caracterização, mais do que a confirmação ou refutação de uma hipótese binária, constitui a contribuição central do trabalho, orientando tanto a aplicação criteriosa do método quanto o desenvolvimento de pesquisa subsequente.

5 CONCLUSÃO

Este trabalho teve por objetivo geral desenvolver e avaliar um modelo de Aprendizado por Reforço, fundamentado no algoritmo Proximal Policy Optimization, para a alocação dinâmica de uma carteira de fundos de índice no mercado brasileiro. O objetivo de construir um modelo funcional, avaliá-lo de forma sistemática sob múltiplos regimes macroeconômicos, cujo comportamento e desempenho foram caracterizados com o rigor que a investigação se propôs a alcançar, foi concluído. Mais do que estabelecer a superioridade incondicional do método, o trabalho delimitou com precisão as condições sob as quais a abordagem é competitiva e aquelas sob as quais não o é.

Os objetivos específicos que sustentaram essa meta foram igualmente atendidos. Procedeu-se à revisão da literatura sobre otimização de portfólio e aprendizado de máquina aplicado a finanças; constituiu-se uma base de dados íntegra e temporalmente alinhada, integrando cotações de cinco ETFs a indicadores macroeconômicos oriundos de fontes oficiais; desenvolveu-se um ambiente de simulação capaz de processar um estado realista, contemplando dados de mercado, variáveis macroeconômicas e a fricção transacional da bolsa; implementou-se e treinou-se o modelo sob a arquitetura PPO, mediante processo iterativo de refinamento; e procedeu-se à análise de sua robustez e de sua capacidade de adaptação em regimes distintos, em confronto quantitativo com estratégias passivas de referência. Cabe destacar que o objetivo de analisar a capacidade de adaptação foi cumprido não pela confirmação de uma hipótese, mas pela sua qualificação criteriosa: a investigação revelou que o modelo não adapta sua alocação ao regime macroeconômico.

A síntese das evidências sustenta uma conclusão dupla. No plano do desempenho, o modelo demonstrou viabilidade prática condicional ao regime: superou a estratégia passiva de Buy & Hold sob juros estáveis, com significância estatística, e preservou retorno positivo na maioria das execuções em todos os cenários, mas cedeu ao Buy & Hold sob o choque pandêmico e foi dominado pelas estratégias de renda fixa sob o aperto monetário. No plano da política, a análise da composição da carteira revelou que essa robustez condicional não decorre de adaptação macroeconômica, já que a alocação média do modelo permanece

aproximadamente invariante entre os regimes, e a variação de seu desempenho origina-se da interação de uma política estável com as condições de cada mercado. O modelo caracteriza-se, portanto, como um alocador tático direcional, que extrai retorno do reposicionamento ativo de alta frequência, e não como um gestor defensivo ou um estrategista macroeconômico.

A contribuição central do trabalho reside menos na proposição de um método superior do que na caracterização precisa e honesta de um sistema de Aprendizado por Reforço aplicado a um problema financeiro realista. Os três elementos fundamentais desenvolvidos para o presente trabalho foram a construção de um ambiente de simulação que incorpora nativamente o contexto macroeconômico e a fricção transacional, a adoção de um protocolo de avaliação multi-seed que, ao expor a dispersão entre inicializações, confere às conclusões grau de confiabilidade superior ao de avaliações fundadas em execução única e a documentação transparente tanto das vantagens condicionais do modelo quanto de suas limitações, a exposição ao risco, a degeneração de execuções sob regimes adversos e a ausência de adaptação macroeconômica, esta última atribuída à falta de normalização das observações.

Tais limitações, longe de invalidarem os resultados, definem o programa de pesquisa subsequente. A normalização das observações, ao tornar o sinal macroeconômico efetivamente acessível à política, permitiria reavaliar de forma justa a hipótese adaptativa. A adoção de funções de recompensa que penalizem explicitamente o risco de queda poderia mitigar a vulnerabilidade observada na eficiência risco-retorno, e a investigação de mecanismos de detecção de degeneração, endereçaria a fragilidade do método sob regimes de transição abrupta. Cada uma dessas direções decorre diretamente de uma limitação aqui documentada, o que confere ao trabalho, ainda em seus limites, valor orientador para a pesquisa futura.

Em última análise, demonstra-se que a aplicação de Aprendizado por Reforço ao rebalanceamento de carteiras no mercado brasileiro é viável e competitiva sob condições específicas, mas está longe de constituir uma solução universal. A contribuição que dele permanece é a compreensão madura de um sistema concreto, de suas capacidades reais e de seus limites igualmente reais, compreensão que constitui fundamento mais sólido para o avanço da área do que a afirmação de uma superioridade que os dados não sustentariam.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ALMAHDI, Saud; YANG, Steve Y. An adaptive portfolio trading system: a risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications*, v. 87, p. 267-279, 2017. Disponível em: <https://doi.org/10.1016/j.eswa.2017.06.023>. Acesso em: 20 set. 2025.

ALMGREN, Robert; CHRISS, Neil. Optimal execution of portfolio transactions. *Journal of Risk*, v. 3, p. 5-40, 2000.

ANG, Andrew. *Asset Management: a systematic approach to factor investing*. Oxford: Oxford University Press, 2014.

ASSOCIAÇÃO BRASILEIRA DAS ENTIDADES DOS MERCADOS FINANCEIRO E DE CAPITAIS (ANBIMA). Índices de Mercado ANBIMA: IRF-M. 2024. Disponível em: https://www.anbima.com.br/data/files/B3/E1/75/0D/63B107109E02B1079B2BA2A8/IRF_M_Lamina.pdf. Acesso em: 16 fev. 2026.

B3 — BRASIL, BOLSA, BALCÃO. Tarifas de Ações e Fundos de Investimento. 2024. Disponível em: https://www.b3.com.br/pt_br/produtos-e-servicos/tarifas/listados-a-vista-e-derivativos/renda-variavel/tarifas-de-acoes-e-fundos-de-investimento/a-vista/. Acesso em: 08 Jan. 2026..

BANCO CENTRAL DO BRASIL (BCB). Taxa Selic. Disponível em: <https://www.bcb.gov.br/controleinflacao/taxaselic>. Acesso em: 9 nov. 2025.

BAUR, Dirk G.; MCDERMOTT, Thomas K. Is gold a safe haven? International evidence. *Journal of Banking & Finance*, v. 34, n. 8, p. 1886-1898, 2010.

BELLMAN, Richard. *Dynamic Programming*. Princeton: Princeton University Press, 1957.

BERNSTEIN, William J. *The Intelligent Asset Allocator: how to build your portfolio to maximize returns and minimize risk*. New York: McGraw-Hill, 2000.

CAMPBELL, John Y.; LO, Andrew W.; MACKINLAY, A. Craig. *The Econometrics of Financial Markets*. Princeton: Princeton University Press, 1997.

CLINTON, Joshua D. Proxy Variable. In: *The SAGE Encyclopedia of Social Science Research Methods*. Thousand Oaks: SAGE Publications, 2011. Disponível em:

<https://methods.sagepub.com/ency/edvol/the-sage-encyclopedia-of-social-science-research-methods/chpt/proxy-variable>. Acesso em: 24 Abril. 2026..

DEMIGUEL, Victor; GARLAPPI, Lorenzo; UPPAL, Raman. Optimal versus naive diversification: how inefficient is the 1/N portfolio strategy? *The Review of Financial Studies*, v. 22, n. 5, p. 1915-1953, 2009. Disponível em: <https://doi.org/10.1093/rfs/hhm075>. Acesso em: 25 Jan. 2026..

FERREIRA, Vitor Augusto. Otimização de Portfólio em Criptoativos utilizando Machine Learning. 2025. Monografia (Graduação em Ciências Econômicas) – Centro Socioeconômico, Universidade Federal de Santa Catarina, Florianópolis, 2025. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/263821/TCC%20-%20Vitor%20Augusto%20Ferreira.pdf?sequence=1&isAllowed=y>. Acesso em: 25 out. 2025.

GASTINEAU, Gary L. *The Exchange-Traded Funds Manual*. 2. ed. Hoboken: John Wiley & Sons, 2010.

GUNJAN, Abhishek; BHATTACHARYYA, Siddhartha. A brief review of portfolio optimization techniques. *Artificial Intelligence Review*, v. 56, p. 3847-3886, 2023. Disponível em: <https://doi.org/10.1007/s10462-022-10273-7>. Acesso em: 20 set. 2025.

HAGSTROM, Robert G. *The Warren Buffett Way*. 2. ed. Hoboken: John Wiley & Sons, 2005.

HASBROUCK, Joel. *Empirical Market Microstructure: the institutions, economics, and econometrics of securities trading*. Oxford: Oxford University Press, 2007.

HENDERSON, Peter; ISLAM, Riashat; BACHMAN, Philip; PINEAU, Joelle; PRECUP, Doina; MEGER, David. Deep Reinforcement Learning that Matters. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 32, n. 1, 2018. Disponível em: <https://arxiv.org/abs/1709.06560>. Acesso em: 28 mar. 2026.

HUANG, Shengyi; DOSSA, Rousslan Fernand Julien; RAFFIN, Antonin; KANERVISTO, Anssi; WANG, Weixun. The 37 Implementation Details of Proximal Policy Optimization. *The ICLR Blog Track*, 2022. Disponível em: <https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/>. Acesso em: 13 Maio. 2026..

HUDSON, Robert S.; GREGORIOU, Andros. Calculating and comparing security returns is harder than you think: a comparison between logarithmic and simple returns. *International Review of Financial Analysis*, v. 38, p. 151-162, 2015.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). Índice Nacional de Preços ao Consumidor Amplo. Disponível em: <https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indice-nacional-de-precos-ao-consumidor-amplo.html>. Acesso em: 9 nov. 2025.

JIANG, Zhengyao; XU, Dixing; LIANG, Jinjun. A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem. arXiv preprint arXiv:1706.10059, 2017. Disponível em: <https://arxiv.org/abs/1706.10059>. Acesso em: 4 out. 2025.

LELAND, Hayne E. Beyond Mean-Variance: performance measurement in a nonsymmetric world. *The Journal of Finance*, v. 54, n. 1, p. 27-36, 1996.

MAGINN, John L.; TUTTLE, Donald L.; PINTO, Jerald E.; MCLEAVEY, Dennis W. *Managing Investment Portfolios: a dynamic process*. 3. ed. New Jersey: John Wiley & Sons, 2007.

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

MASTERS, Seth J. Rebalancing. *The Journal of Portfolio Management*, v. 29, n. 3, p. 52-57, 2003.

MITCHELL, Tom M. *Machine Learning*. New York: McGraw-Hill, 1997.

MNIH, Volodymyr; KAVUKCUOGLU, Koray; SILVER, David et al. Human-level control through deep reinforcement learning. *Nature*, v. 518, n. 7540, p. 529-533, 2015. Disponível em: <https://doi.org/10.1038/nature14236>. Acesso em: 03 Abril. 2026..

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS (PUC MINAS). ANÁLISE QUANTITATIVA DAS CARTEIRAS AUTOMATIZADAS DA XP INVESTIMENTOS. 2023. Disponível em: <http://bib.pucminas.br:8080/pergamumweb/vinculos/0000b9/0000b980.pdf>. Acesso em: 12 Jan. 2026..

RAFFIN, Antonin; HILL, Ashley; GLEAVE, Adam; KANERVISTO, Anssi; ERNESTUS, Maximilian; DORMANN, Noah. Stable-Baselines3: reliable reinforcement learning implementations. *Journal of Machine Learning Research*, v. 22, n. 268, p. 1-8, 2021. Disponível em: <https://jmlr.org/papers/v22/20-1364.html>. Acesso em: 19 Dez. 2025.

SCHULMAN, John; LEVINE, Sergey; ABBEEL, Pieter; JORDAN, Michael; MORITZ, Philipp. Trust Region Policy Optimization. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. Lille: PMLR, 2015. v. 37, p. 1889-1897. Disponível em: <https://proceedings.mlr.press/v37/schulman15.html>. Acesso em: 01 Fev. 2026..

SCHULMAN, John; WOLSKI, Filip; DHARIWAL, Prafulla; RADFORD, Alec; KLIMOV, Oleg. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. Disponível em: <https://arxiv.org/abs/1707.06347>. Acesso em: 29 Dez. 2025..

SEN, Jaydip. A Comparative Study on the Sharpe Ratio, Sortino Ratio, and Calmar Ratio in Portfolio Optimization. Preprint (ResearchGate), 2022. Disponível em: https://www.researchgate.net/profile/Jaydip-Sen-3/publication/366517929_A_Comparative_Study_on_the_Sharpe_Ratio_Sortino_Ratio_and_Calmar_Ratio_in_Portfolio_Optimization. Acesso em: 23 out. 2025.

SENDACZ, Guilherme T. Utilização do Dólar Americano como Hedge e/ou Porto Seguro para a Carteira do Investidor Brasileiro. 2015. Trabalho de Conclusão de Curso (Graduação em Administração) – Insper Instituto de Ensino e Pesquisa, São Paulo, 2015.

SUTTON, Richard S.; BARTO, Andrew G. Reinforcement Learning: an introduction. 2. ed. Cambridge: MIT Press, 2018.

SWENSEN, David F. Unconventional Success: a fundamental approach to personal investment. New York: Free Press, 2005.

THEODORO, Tiago Piassum. Bolsa de Valores e as Variáveis Macroeconômicas do Brasil: uma análise utilizando modelos VAR. 2018. Monografia (Graduação em Ciências Econômicas) – Universidade Federal do Rio Grande, Rio Grande, 2018. Disponível em: <https://economia.furg.br/images/banners/Monografias/20181/tiagopiassum.pdf>. Acesso em: 1 nov. 2025.

UPPALADINNI, Shiva et al. Machine Learning for Real-Time Portfolio Rebalancing: a novel approach to financial optimization. International Journal for Research in Applied Science & Engineering Technology (IJRASET), v. 12, n. 10, out. 2024. Disponível em: [https://www.researchgate.net/profile/Shiva-](https://www.researchgate.net/profile/Shiva-Uppaladinni/publication/384633823_Machine_Learning_for_Real-Time_Portfolio_Rebalancing_A_Novel_Approach_to_Financial_Optimization)

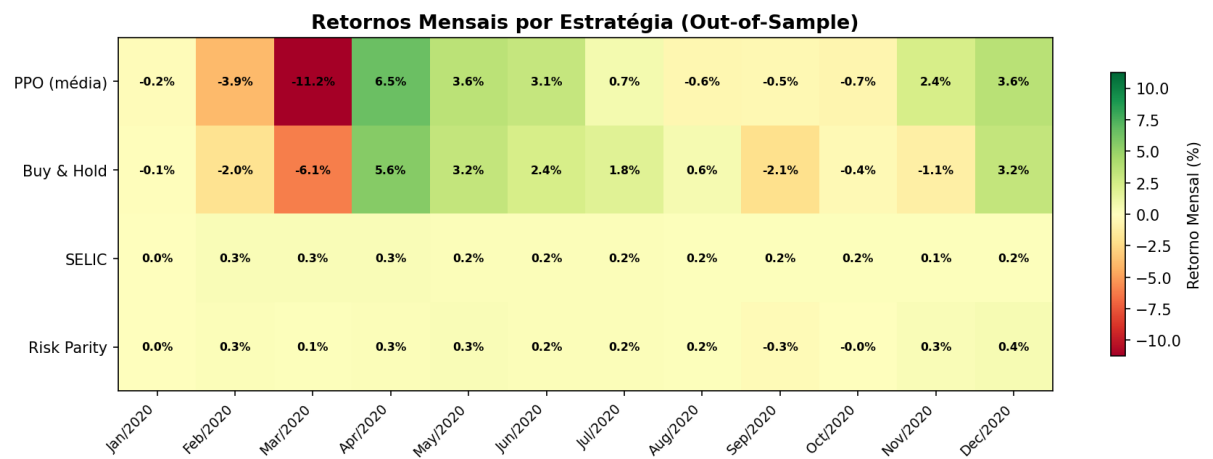
[Uppaladinni/publication/384633823_Machine_Learning_for_Real-Time_Portfolio_Rebalancing_A_Novel_Approach_to_Financial_Optimization](https://www.researchgate.net/profile/Shiva-Uppaladinni/publication/384633823_Machine_Learning_for_Real-Time_Portfolio_Rebalancing_A_Novel_Approach_to_Financial_Optimization). Acesso em: 26 set. 2025.

VANGUARD RESEARCH. Best practices for portfolio rebalancing. Valley Forge: Vanguard White Paper, 2019.

WOOLDRIDGE, Jeffrey M. Introductory Econometrics: a modern approach. 6. ed. Boston: Cengage Learning, 2016.

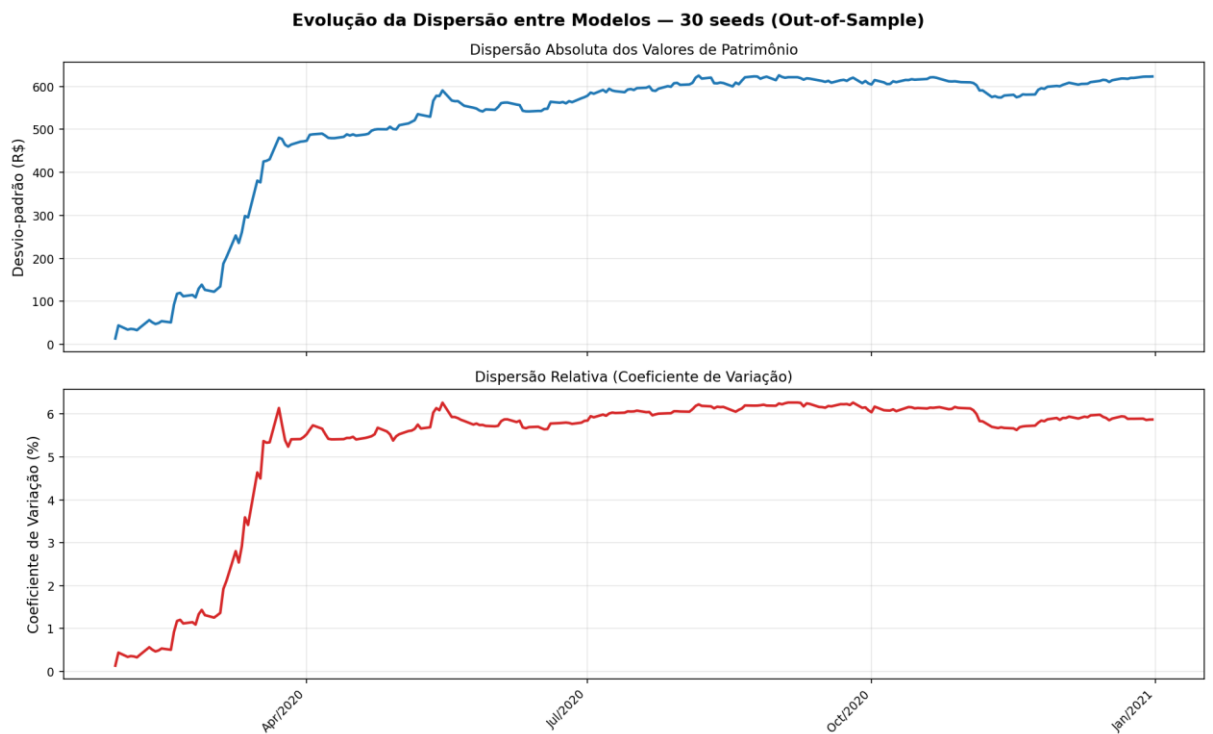
APÊNDICE A – FIGURAS COMPLEMENTARES DOS CENÁRIOS DE ESTRESSE

Figura A.1 – Mapa de calor dos retornos mensais - cenário COVID-19 (2020)



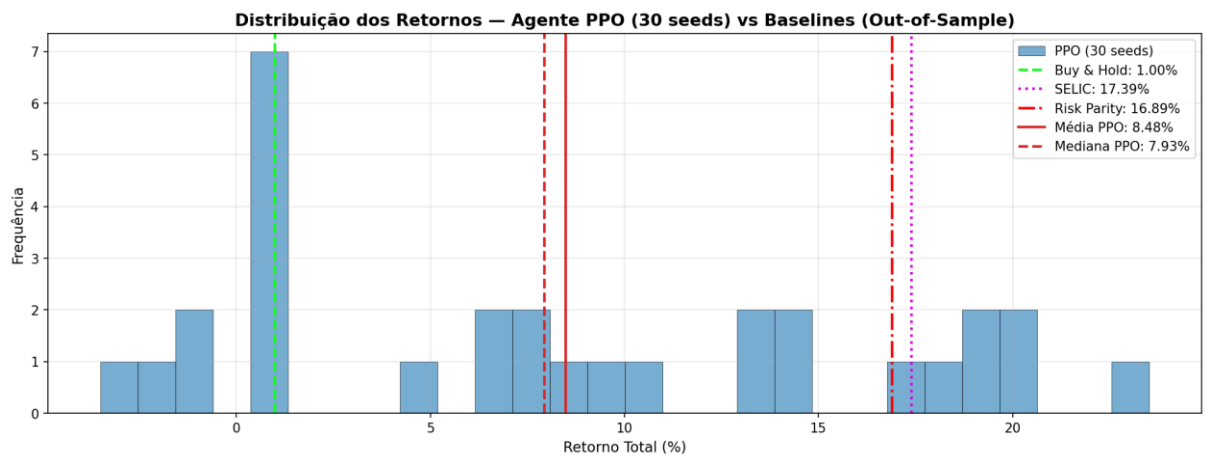
Fonte: Dados da pesquisa (2026)

Figura A.2 – Evolução da dispersão entre execuções - cenário COVID-19 (2020)



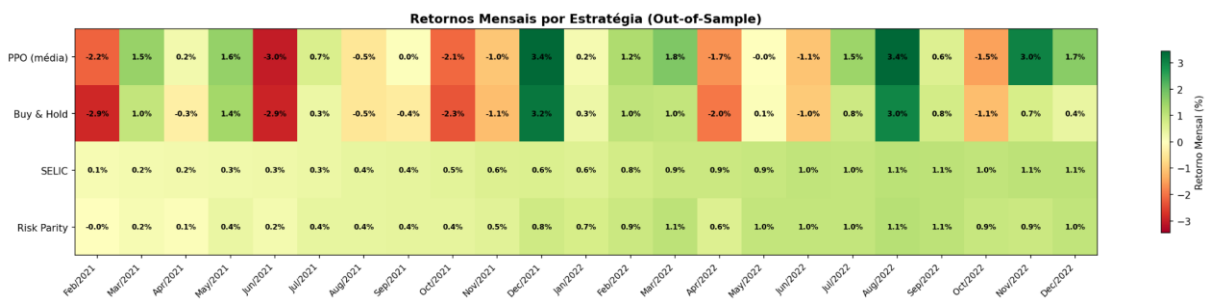
Fonte: Dados da pesquisa (2026)

Figura A.3 – Histograma dos retornos das execuções - cenário de expansão da SELIC (2021–2022)



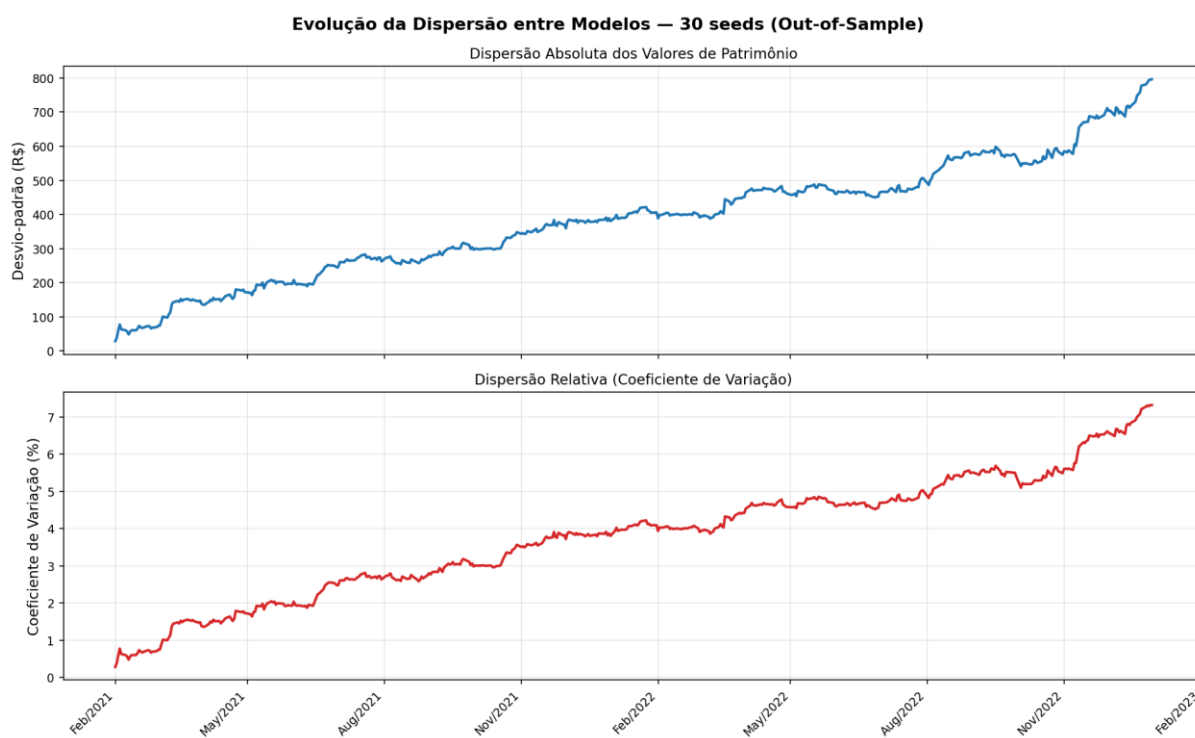
Fonte: Dados da pesquisa (2025)

Figura A.4 – Mapa de calor dos retornos mensais - cenário de expansão da SELIC (2021–2022)



Fonte: Dados da pesquisa (2025)

Figura A.5 – Evolução da dispersão entre execuções - cenário de expansão da SELIC (2021–2022)



Fonte: Dados da pesquisa (2025)

APÊNDICE B – Código Fonte do Modelo desenvolvido

- Link de exposição do código fonte:

<https://codigos.ufsc.br/guilherme.cardoso.oliveira/tcc-portfolio-optimization-using-rl.git>

APÊNDICE C – Artigo

APLICAÇÃO DE APRENDIZADO DE MÁQUINA PARA REBALANCEAMENTO DE CARTEIRA DE INVESTIMENTOS

Guilherme Cardoso de Oliveira ¹

¹Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC)
88.040-900 – Florianópolis – SC – Brasil
guicardosodeoliveira@gmail.com

Abstract. *This paper evaluates a Proximal Policy Optimization (PPO) agent for the dynamic rebalancing of a five-ETF portfolio in the Brazilian market, within a simulation environment that embeds macroeconomic indicators (SELIC and IPCA) into the state space and models transaction friction natively. The reward is the cost-net log-return. Under a multi-seed protocol, performance was assessed out-of-sample across three macroeconomic regimes and compared against Buy & Hold, SELIC and inverse-volatility baselines. Results reveal regime-conditional robustness: the agent beats the passive baseline under stable rates, yields to it under an acute shock, and is dominated by fixed income under monetary tightening. Portfolio composition remains nearly invariant across regimes, characterizing the agent as a tactical directional allocator rather than a macro-aware manager.*

Resumo. *Este artigo avalia um agente Proximal Policy Optimization (PPO) para o rebalanceamento dinâmico de uma carteira de cinco ETFs do mercado brasileiro, em um ambiente de simulação que integra indicadores macroeconômicos (SELIC e IPCA) ao espaço de estados e modela nativamente a fricção transacional. A recompensa adotada é o log-retorno líquido de custos. Sob protocolo multi-seed, o desempenho foi aferido out-of-sample em três regimes macroeconômicos e confrontado com as estratégias Buy & Hold, SELIC e paridade de risco por volatilidade inversa. Os resultados evidenciam robustez condicional ao regime: o agente supera a estratégia passiva sob juros estáveis, cede-lhe sob choque agudo e é dominado pela renda fixa sob aperto monetário. A composição da carteira permanece aproximadamente invariante entre regimes, caracterizando o agente como alocador tático direcional, e não como gestor macroeconômico.*

1. Introdução

A gestão de carteiras em mercados emergentes é exercício de adaptação contínua a regimes macroeconômicos instáveis, e não mera seleção de ativos rentáveis. A Teoria Moderna do Portfólio [Markowitz 1952], marco fundacional da diversificação, apoia-se em parâmetros estáticos, vetores de retornos esperados e matrizes de covariância estimados sobre dados pretéritos, o que a torna sensível a erros de estimação e pouco resiliente a rupturas abruptas de política monetária [Uppaladinni et al. 2024]. A própria superioridade empírica dessas formulações sobre regras ingênuas de alocação é objeto de contestação consolidada na literatura [DeMiguel et al. 2009], o que motiva a investigação de métodos capazes de reconfigurar a alocação de forma endógena ao longo do tempo.

O Aprendizado por Reforço Profundo oferece formulação natural para esse problema, ao modelar a alocação como decisão sequencial sob incerteza, na qual a política é otimizada diretamente sobre o objetivo econômico, sem etapa intermediária de previsão de preços [Sutton and Barto 2018, Mnih et al. 2015]. Trabalhos seminais na área [Jiang et al. 2017] demonstraram a viabilidade da abordagem, mas o campo permanece marcado por avaliações conduzidas sobre execução única e ambientes que abstraem a fricção transacional, práticas cuja fragilidade metodológica é amplamente documentada [Henderson et al. 2018].

Este artigo, derivado de um trabalho de conclusão de curso, investiga a viabilidade e a robustez de um agente PPO [Schulman et al. 2017] para o rebalanceamento dinâmico de uma carteira de cinco fundos de índice brasileiros, em um ambiente que integra indicadores macroeconômicos ao espaço de estados e incorpora nativamente os custos de negociação da bolsa. A hipótese que norteou o desenho experimental é a de que o acesso ao contexto macroeconômico conferiria ao agente capacidade de adaptar sua alocação ao regime vigente. A contribuição do trabalho é dupla e reside menos na proposição de um método superior do que na caracterização rigorosa de um sistema concreto: de um lado, documenta-se, sob protocolo multi-seed e em três regimes distintos, uma robustez estritamente condicional ao contexto macroeconômico; de outro, demonstra-se que essa condicionalidade não decorre de adaptação deliberada da política, hipótese que a evidência refuta, mas da interação entre uma política aproximadamente invariante e as condições idiossincráticas de cada mercado.

2. Trabalhos Relacionados

A literatura de otimização de portfólio evoluiu da formulação média-variância clássica para abordagens que incorporam aprendizado de máquina, movimento sintetizado por Gunjan e Bhattacharyya (2023) em revisão que classifica as técnicas segundo o grau de dependência de premissas distribucionais. No âmbito do Aprendizado por Reforço, a arquitetura EIIE proposta por Jiang et al. (2017) constitui referência metodológica central, ao formalizar o mapeamento entre a saída da rede e o vetor de pesos e ao embutir o custo de transação no sinal de recompensa, decisões de desenho aqui adotadas. Almahdi e Yang (2017), por sua vez, propõem funções de recompensa ajustadas ao risco que penalizam explicitamente o drawdown esperado, formulação deliberadamente não adotada neste trabalho, cuja recompensa minimalista delega a contenção de risco à estrutura do ambiente.

A lacuna que este trabalho endereça situa-se na conjunção de três elementos raramente presentes de forma simultânea na literatura aplicada ao mercado brasileiro: a incorporação de variáveis macroeconômicas ao espaço de estados como mecanismo de contextualização de regime, a modelagem realista da fricção transacional segundo a estrutura de tarifas efetivamente vigente [B3 2024, Almgren and Chriss 2000], e a adoção de um protocolo de avaliação multi-seed que expõe a dispersão entre inicializações, em conformidade com as recomendações de reprodutibilidade formuladas por Henderson et al. (2018).

3. Metodologia

O percurso metodológico organiza-se em quatro etapas encadeadas: a coleta e o tratamento dos dados, a construção do ambiente de simulação, o treinamento do agente sob refinamento iterativo e o protocolo de avaliação. O universo de investimento compreende cinco ETFs negociados na B3, selecionados segundo critérios de representatividade de classe, liquidez e profundidade histórica: BOVA11 (renda variável), IMAB11 (títulos indexados à inflação), IRFM11 (títulos prefixados), XFIX11 (fundos imobiliários) e GOLD11 (ouro). O problema de rebalanceamento foi modelado como um Processo de Decisão de Markov, ressaltando-se que se trata de uma modelagem, e não de propriedade intrínseca do mercado, que é apenas parcialmente observável.

3.1. Coleta e tratamento dos dados

A base cobre o período de 2015 a 2024 em frequência diária e integra três fontes: as cotações dos ETFs, obtidas por interface programática de dados de mercado, e as séries macroeconômicas da taxa SELIC e do IPCA, extraídas do Sistema Gerenciador de Séries Temporais do Banco Central do Brasil. A juventude relativa dos fundos de índice brasileiros impõe um obstáculo à extensão da base, condição indispensável para expor o agente a múltiplos regimes. Recorreu-se, por isso, à construção de séries proxy para os ativos de histórico incompleto, procedimento admissível quando a série substituta exibe forte correlação com a variável-alvo no intervalo de sobreposição [Wooldridge 2016, Clinton 2011].

Três mecanismos foram empregados conforme a natureza da indisponibilidade. O XFIX11 foi substituído integralmente pelo KNRI11, fundo imobiliário de perfil semelhante e histórico mais longo ($r = 0,9486$). O GOLD11 foi reconstruído por série sintética, mediante o produto entre a cotação do ETF internacional GLD e a taxa de câmbio USD/BRL ($r = 0,9734$). O IMAB11 e o IRFM11 foram obtidos por costura temporal entre os índices oficiais IMA-B e IRF-M, divulgados pelo Banco Central, e os respectivos ETFs, com reescalonamento de nível no ponto de junção para preservar a continuidade da série ($r = 0,9862$ e $r = 0,9033$). Todas as aderências foram validadas pelo coeficiente de correlação de Pearson sobre o período de sobreposição, e todas superam 0,90.

O conjunto integrado foi reindexado sobre um calendário regular de dias úteis, a fim de harmonizar os calendários heterogêneos das fontes, e as observações ausentes foram preenchidas por propagação do último valor válido, sem qualquer interpolação que recorresse a dados futuros. A prevenção do viés de antecipação exigiu tratamento específico do IPCA, cujo valor de competência de um mês só é divulgado semanas após o encerramento do período de referência: a série foi deslocada em um mês, de modo que o agente receba a inflação de janeiro somente em fevereiro, simulando a defasagem entre a competência do índice e sua disponibilização pública. Sobre as séries de preços derivaram-se, por ativo, o log-retorno diário e três indicadores móveis, a volatilidade, o índice de Sharpe e o índice de Sortino, que compõem, ao lado dos indicadores macroeconômicos, o vetor de atributos do estado.

3.2. Construção do ambiente de simulação

O ambiente expõe ao agente uma observação composta por dois blocos. O primeiro é uma matriz 21×22 , na qual a primeira dimensão corresponde à janela deslizante dos vinte e um pregões mais recentes e a segunda reúne as vinte variáveis derivadas dos cinco ativos, acrescidas da taxa SELIC diária e do IPCA. O segundo é o vetor de pesos correntes da carteira, cuja inclusão é condição necessária para que a política possa antecipar o custo do rebalanceamento implicado por uma nova alocação. Registre-se, pela centralidade que o ponto assumirá na discussão dos resultados, que os atributos do estado não foram submetidos a qualquer normalização, sendo alimentados à rede em suas escalas originais.

O espaço de ações é contínuo, definido sobre uma caixa em $\mathbb{R}^5$. A saída do Ator não corresponde diretamente a pesos, mas a logits convertidos em uma distribuição válida pela função softmax e submetidos, em seguida, a um procedimento iterativo de redistribuição que trunca em quarenta por cento o peso de qualquer ativo e realoca o excedente proporcionalmente aos demais. Essa separação entre o espaço bruto de saída da rede e o simplex de pesos permite que a otimização opere em espaço euclidiano irrestrito, deixando as restrições de não negatividade, soma unitária e concentração máxima a cargo do mapeamento [Jiang et al. 2017].

A recompensa a cada pregão é o log-retorno líquido de custos do portfólio, definido como o produto interno entre o vetor de pesos efetivamente em vigor e o vetor de log-retornos dos ativos, acrescido do custo logarítmico da transação:

$$r_t = w_t^{exec} \cdot \rho_t + c_t \quad (1)$$

A adoção do log-retorno assegura que a soma das recompensas, objetivo de otimização do PPO, corresponda exatamente ao crescimento acumulado do patrimônio [Campbell et al. 1997], ao passo que a dedução imediata do custo faz com que o agente internalize o dilema entre corrigir o desvio da carteira e preservar capital contra o atrito operacional, sem necessidade de regras rígidas [Jiang et al. 2017]. A recompensa não penaliza explicitamente métricas de risco, decisão deliberada que delega a contenção de risco à estrutura do ambiente e cujas consequências são examinadas na Seção 5.

O acionamento do rebalanceamento e a cobrança do custo são governados pelo turnover τ implicado pela ação, medido pelo maior desvio absoluto individual entre o vetor de pesos-alvo e o vetor de pesos vigentes. A operação só é executada quando esse desvio supera o limiar τ_{min} , mecanismo que materializa computacionalmente a zona de não-transação teorizada por Leland (1996) e a prática institucional de bandas de tolerância por ativo [Masters 2003, Vanguard Research 2019]:

$$c_t = \ln(1 - \tau_t \cdot f), \text{ se } \tau_t > \tau_{min}; \quad c_t = 0, \text{ caso contrário} \quad (2)$$

O limiar foi fixado em cinco pontos percentuais, em consonância com a recomendação empírica da Vanguard Research (2019), e a taxa proporcional em $f = 0,03\%$, calibrada segundo a estrutura tarifária vigente para ETFs de alta liquidez [B3 2024]. Essa formulação captura o componente linear do modelo de custos de Almgren e Chriss (2000), abstraindo os termos não lineares de impacto de mercado, desprezáveis para o capital inicial de R\$ 10.000,00 adotado na simulação.

3.3. Treinamento e refinamento iterativo

O agente foi implementado com o algoritmo PPO da biblioteca Stable-Baselines3 [Raffin et al. 2021], sob política MultiInputPolicy, adequada ao espaço de observação composto. A configuração final não é fruto de escolha arbitrária, mas de um processo iterativo de refinamento conduzido em lotes sucessivos, monitorado pelas métricas internas de treinamento. A configuração inicial exibiu instabilidade na dinâmica de otimização, diagnosticada pela inspeção conjunta da fração de recorte, da divergência de Kullback-Leibler aproximada e da variância explicada, cujos intervalos saudáveis são documentados na literatura de implementação do algoritmo [Huang et al. 2022]. O ajuste dos hiperparâmetros do otimizador estabilizou a convergência, mas fez emergir hiperatividade transacional, endereçada pela introdução do limiar de rebalanceamento e da restrição de concentração descritos na subseção anterior. Testou-se, ademais, uma variante da recompensa que subtraía a SELIC diária do retorno do portfólio, sob a premissa de que penalizar o custo de oportunidade da renda fixa induziria postura mais conservadora; a formulação prejudicou a convergência e foi abandonada em favor do log-retorno líquido puro.

Durante o treinamento, um mecanismo de avaliação periódica sobre o conjunto de validação seleciona, a cada dez mil timesteps, o melhor modelo segundo o desempenho fora da amostra de treino, de modo que o modelo submetido ao teste não é o do último passo de otimização, mas o de melhor desempenho em validação. A Tabela 1 consolida a configuração final, empregada sem alteração em todos os cenários de avaliação.

Tabela 1. Configuração final do agente e do ambiente

Parâmetro	Valor	Componente
Algoritmo / política	PPO / MultiInputPolicy	Modelo
Taxa de aprendizado	1×10^{-4}	Treinamento
Tamanho do rollout	4096	Treinamento
Tamanho de minibatch	64	Treinamento
Fator de desconto (γ)	0,99	Treinamento
Timesteps por execução	1.000.000	Treinamento
Frequência de avaliação	10.000 timesteps	Treinamento
Janela de observação	21 pregões	Ambiente

Parâmetro	Valor	Componente
Capital inicial	R\$ 10.000,00	Ambiente
Taxa de transação (f)	0,03%	Ambiente
Limiar de rebalanceamento 5 p.p. (tmin)		Ambiente
Concentração máxima por ativo	40%	Ambiente
Função de recompensa	Log-retorno líquido	Ambiente

3.4. Protocolo de avaliação

Dada a sensibilidade documentada do método à inicialização aleatória [Henderson et al. 2018], o desempenho aferido em execução única não constitui evidência confiável da qualidade da configuração, pois confunde a competência da política com a sorte da semente. Adotou-se, por isso, protocolo multi-seed: o cenário principal foi avaliado sobre duzentas e uma execuções independentes do procedimento completo de treinamento e o de cada cenário de estresse sobre trinta, sendo todos os resultados reportados como agregações estatísticas sobre esse conjunto. Os resultados são reportados por média e desvio-padrão, complementados pela mediana sempre que esta se justifique como descritor mais representativo da tendência central, dado que sementes ocasionalmente convergem para políticas degeneradas e deslocam a média de forma desproporcional.

A avaliação abrange três regimes macroeconômicos, cada qual com divisão temporal estrita e independente entre treino, validação e teste. O cenário principal compreende o biênio 2023–2024, caracterizado por juros elevados e comparativamente estáveis, com validação sobre 2022 e treino sobre o período anterior. O primeiro cenário de estresse corresponde ao choque pandêmico de 2020, marcado por queda abrupta seguida de recuperação em V sob SELIC em mínima histórica, com validação sobre 2019. O segundo corresponde ao ciclo de aperto monetário de 2021–2022, no qual a taxa básica ascendeu de 2% a 13,75%, com validação sobre 2020. Em todos os casos, o período de teste é integralmente posterior aos dados de treinamento e validação, e a configuração do agente permanece inalterada, de modo que as diferenças observadas sejam atribuíveis às características de cada regime, e não a variações do procedimento experimental.

O desempenho é confrontado com três estratégias de referência, executadas sobre o mesmo intervalo out-of-sample e sob a mesma estrutura de custos: o Buy & Hold equiponderado, que representa a gestão inteiramente passiva e constitui o piso de valor agregado que qualquer estratégia ativa precisa superar [Ang 2014]; a aplicação integral à SELIC capitalizada diariamente, que representa o retorno

obtido sem assunção de risco de mercado; e uma estratégia de paridade de risco, cujos pesos são inversamente proporcionais às volatilidades estimadas dos ativos e recalculados mensalmente. As métricas empregadas articulam-se em três dimensões: retorno total e anualizado; risco, pela volatilidade anualizada e pelo drawdown máximo; e eficiência risco-retorno, pelos índices de Sharpe, Sortino e Calmar. O atrito operacional é quantificado pelo número de rebalanceamentos e pelo custo cumulativo de transação.

4. Resultados

Todos os resultados referem-se a períodos out-of-sample. A inferência foi conduzida por testes t de uma amostra, escolha ditada pela assimetria das quantidades comparadas: o agente produz uma distribuição de retornos, uma observação por semente, ao passo que cada baseline é determinística sobre o conjunto de teste e resulta em um valor único. Cumpre delimitar o alcance dessa inferência, uma vez que a variabilidade entre execuções decorre exclusivamente da inicialização aleatória, e não de múltiplas realizações de mercado, de modo que os testes aferem se o resultado esperado do procedimento difere das referências sobre um único caminho histórico, não constituindo medida de generalização a outras trajetórias.

4.1. Cenário principal (2023–2024)

Tabela 2. Estatísticas descritivas do cenário principal (2023–2024)

Métrica	PPO (n = 201)	Buy & Hold	SELIC	Risk Parity
Retorno total (%)	25,59	22,32	25,12	14,71
Retorno anualizado (%)	12,17	10,71	11,98	7,18
Volatilidade anualizada (%)	7,69	6,66	0,08	4,64
Máximo drawdown (%)	–5,26	–3,62	0,00	–6,33
Índice de Sharpe	0,044	–0,147	—	–0,936
Índice de Sortino	0,071	–0,200	—	–1,178
Índice de Calmar	2,49	2,96	—	1,13

Métrica	PPO (n = 201)	Buy & Hold	SELIC	Risk Parity
Rebalanceamentos (média)	372,7	—	—	—
Custo total (R\$)	200,68	—	—	—

O agente registra retorno total médio de 25,59%, com mediana de 25,64% e desvio-padrão de 5,42 pontos percentuais entre execuções, sem que qualquer semente incorra em prejuízo. O valor supera o Buy & Hold e domina amplamente a paridade de risco, aproximando-se do rendimento acumulado da SELIC. O índice de Sharpe quase nulo (0,044) qualifica esse desempenho: conquanto o retorno absoluto seja expressivo, o excedente sobre o ativo livre de risco é apenas marginal, de sorte que o agente assume risco direcional de mercado para auferir rendimento estatisticamente equivalente ao da renda fixa. O índice de Calmar inferior ao da estratégia passiva (2,49 contra 2,96) antecipa a vulnerabilidade estrutural discutida na Seção 5. No plano operacional, cada execução realiza em média 372,7 rebalanceamentos, ao custo cumulativo médio de R\$ 200,68, aproximadamente dois por cento do capital inicial ao longo do biênio, integralmente deduzido dos retornos reportados.

Tabela 3. Análise inferencial do desempenho frente às estratégias de referência (cenário principal)

Comparação	Diferença média (p.p.)	Estatística t	p-valor	IC 95% (p.p.)	Superação (%)
PPO vs. Buy & Hold	+3,27	+8,54	$\approx 3 \times 10^{-15}$	[2,51; 4,02]	69,7
PPO vs. SELIC	+0,47	+1,22	0,22	[-0,29; 1,22]	54,2
PPO vs. Risk Parity	+10,88	+28,42	$\approx 4 \times 10^{-72}$	[10,12; 11,63]	98,5

A superação do Buy & Hold é estatisticamente robusta, com diferença média de 3,27 pontos percentuais, ao passo que a confrontação com a SELIC não permite rejeitar a hipótese de igualdade, configurando empate técnico coerente com o Sharpe quase nulo. Registre-se que a magnitude da significância é amplificada pelo tamanho amostral: o p-valor diminuto atesta a consistência do resultado entre inicializações, e não a dimensão econômica da vantagem, adequadamente representada pela diferença de médias. A Figura 1 apresenta a curva de capital multi-seed do cenário.

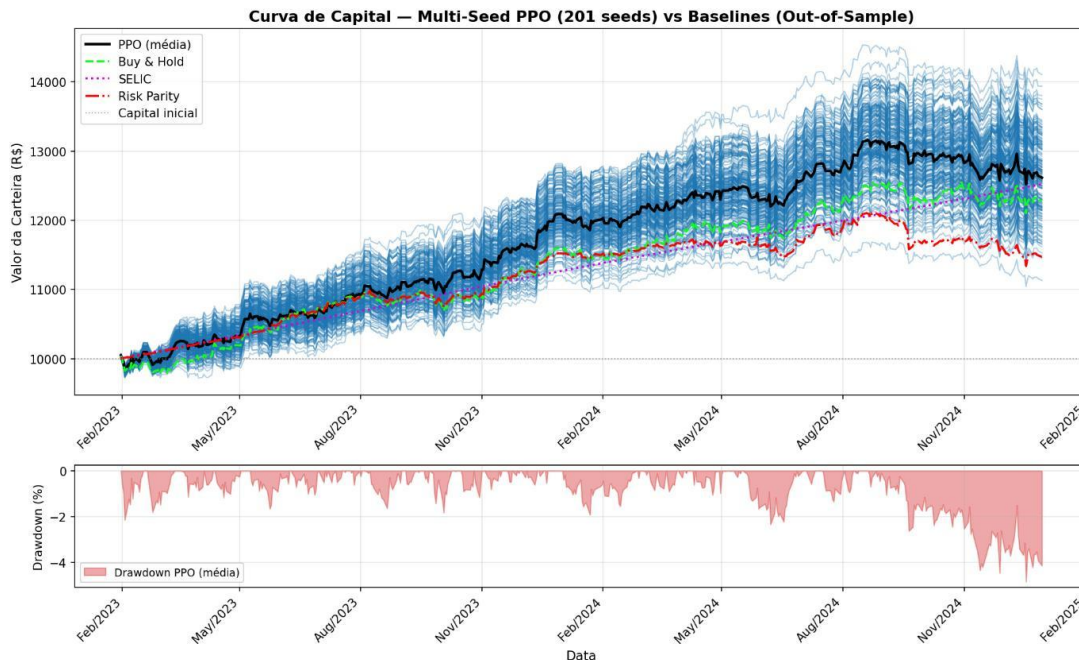


Figura 1. Curva de capital multi-seed e estratégias de referência – cenário principal (2023–2024)

4.2. Choque pandêmico (2020)

Sob o choque agudo de 2020, o agente obtém retorno médio de 5,90%, ficando atrás do Buy & Hold (9,92%) por margem significativa ($t = -3,53$; $p = 0,0014$), com drawdown máximo médio de $-22,83\%$, mais profundo que o da própria estratégia passiva ($-13,62\%$). Preserva, contudo, ampla superioridade sobre as alternativas de renda fixa e encerra o período com retorno positivo em 86,7% das execuções. O comportamento dos ativos esclarece o resultado: o ouro disparou no período, ao passo que a renda variável e os fundos imobiliários acumularam perdas pronunciadas, de sorte que a carteira diversificada do agente foi simultaneamente penalizada pela exposição ao risco e amparada pela posição em ouro, sem capturar a recuperação com a eficiência da carteira equiponderada. Emerge, portanto, o perfil de um alocador direcional que não atua como gestor defensivo em episódios de queda abrupta.

4.3. Aperto monetário (2021–2022)

No ciclo de elevação da SELIC, o agente registra retorno médio de 8,48% e Sharpe negativo ($-0,416$), consequência direta de um rendimento inferior ao do ativo livre de risco em regime de juros muito elevados. Supera o Buy & Hold, que praticamente estagnou (1,00%; $t = +5,06$), mas é dominado de modo significativo tanto pela SELIC (17,39%; $t = -6,02$) quanto pela paridade de risco (16,89%; $t = -5,69$), estratégias que se beneficiaram dos únicos ativos rentáveis do período, os títulos de renda fixa. O cenário revela, ademais, bimodalidade severa na distribuição dos desempenhos, ilustrada na Figura 2: um terço das execuções degenera em políticas inertes, com menos de cem rebalanceamentos e retorno médio de apenas 0,15%, ao passo que as execuções ativas rendem em média 12,65%, com correlação forte entre atividade transacional e retorno ($r = 0,738$; $p \approx 3 \times 10^{-6}$). Essa bimodalidade atenua a adequação da suposição de normalidade subjacente ao teste t , razão pela qual as conclusões deste cenário extraem sua sustentação principal da análise de distribuição.

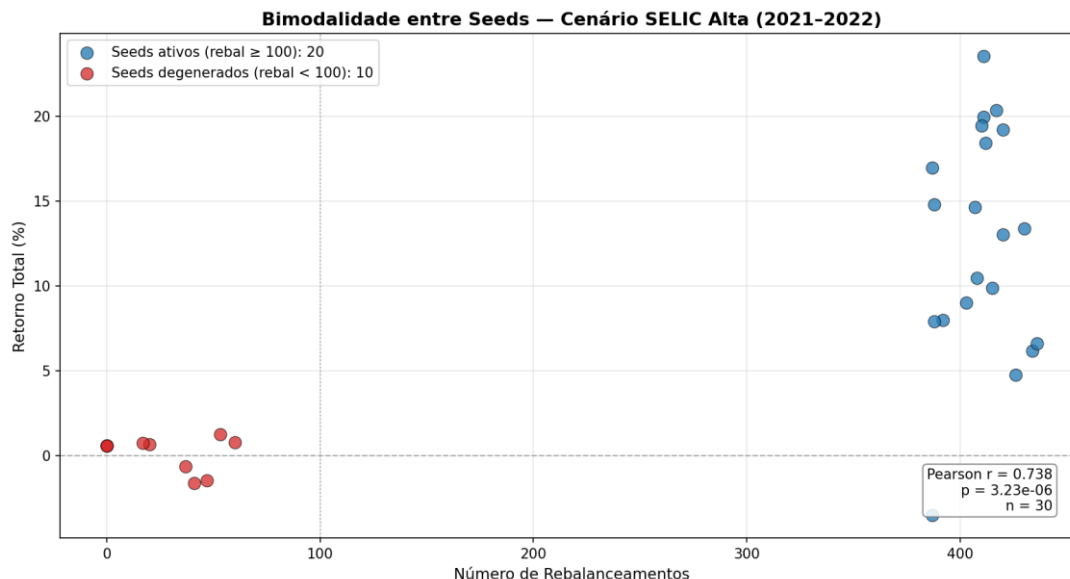


Figura 2. Retorno por execução versus número de rebalanceamentos – cenário de aperto monetário (2021–2022)

O contraste com a degeneração observada na pandemia é instrutivo. Lá, a inação era benigna, pois a recuperação em V recompensava a simples permanência no mercado; aqui, como a alta dos juros exigia migração ativa para a renda fixa, a mesma inação condena a execução à estagnação.

4.4. Síntese comparativa entre regimes

Tabela 4. Síntese comparativa do desempenho entre regimes macroeconômicos

Métrica	Principal (n = 201)	COVID-19 (n = 30)	Aperto monetário (n = 30)
Retorno PPO (%)	25,59	5,90	8,48
Retorno Buy & Hold (%)	22,32	9,92	1,00
Retorno SELIC (%)	25,12	2,51	17,39
Retorno Risk Parity (%)	14,71	2,08	16,89
Índice de Sharpe (PPO)	0,044	0,270	−0,416
Índice de Calmar (PPO)	2,49	0,33	0,62

Métrica	Principal (n = 201)	COVID-19 (n = 30)	Aperto monetário (n = 30)
Índice de Calmar (B&H)	2,96	0,77	0,08
Execuções positivas (%)	100,0	86,7	86,7
Execuções degeneradas (%)	1,0	23,3	33,3

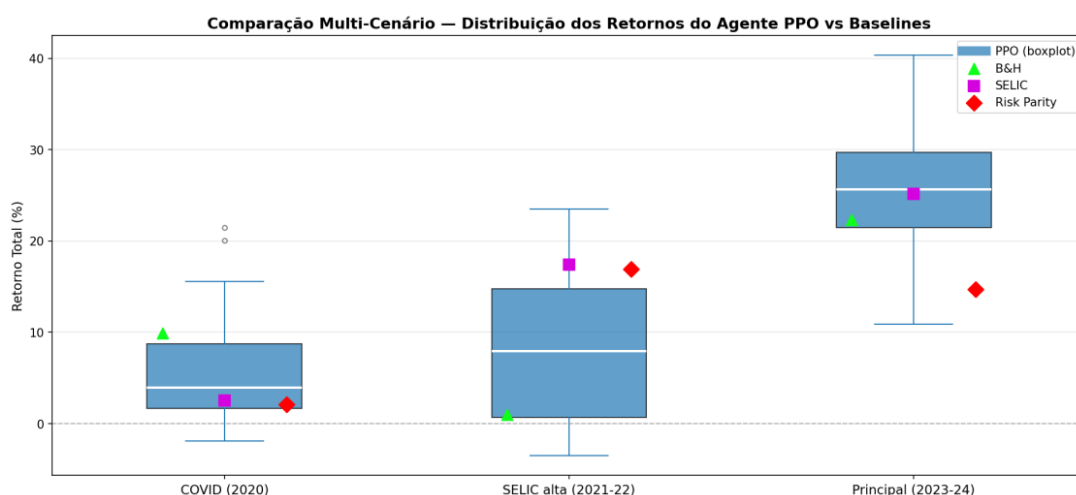


Figura 3. Distribuição dos retornos do agente nos três regimes (diagramas de caixa)

O quadro consolidado da Tabela 4 evidencia desempenho fortemente dependente do contexto, com a identidade do melhor referente variando conforme o regime: a SELIC e a paridade de risco, dominadas sob juros estáveis, tornam-se as estratégias superiores sob aperto monetário. Resta confrontar diretamente a hipótese que motivou a incorporação das variáveis macroeconômicas ao estado. A composição média da carteira, agregada sobre sementes e período, mantém-se próxima da equiponderação nos três regimes, com os pesos dos cinco ativos confinados a uma faixa estreita em torno de vinte por cento e variação inter-regime não superior a sete pontos percentuais. O único movimento dotado de coerência econômica, a redução gradual da exposição à renda variável de 25,2% para 20,1% sob aperto, é modesto e vem acompanhado de aumento da posição em fundos imobiliários no mesmo regime, movimento que contraria a sensibilidade negativa desses ativos aos juros e descaracteriza qualquer padrão adaptativo consistente. Repetida a análise apenas sobre as sementes ativas, para descartar artefato das execuções degeneradas, o padrão persiste com diferenças inferiores a um ponto percentual. Conclui-se que o agente não adapta sua alocação ao regime macroeconômico.

5. Discussão

A hipótese que norteou o trabalho confirma-se de forma qualificada. O agente demonstrou viabilidade prática inequívoca, encerrando todos os regimes com retorno positivo na maioria das execuções e superando a estratégia passiva em dois dos três cenários com significância estatística, mas sua robustez é condicional, e não absoluta. A contribuição empírica reside, portanto, menos em estabelecer a superioridade do método do que em delimitar com precisão as condições sob as quais ele é competitivo.

A primeira limitação, transversal aos cenários, concerne à eficiência risco-retorno: o agente entrega retorno mediante exposição proporcionalmente maior ao risco de cauda. Trata-se de característica estrutural, e não de artefato, pois a recompensa adotada premia o ganho sem penalizar explicitamente os drawdowns, de modo que a política não dispõe de incentivo para limitar perdas de pico a vale, vulnerabilidade cuja mitigação passa por formulações de recompensa ajustadas ao risco [Almahdi and Yang 2017]. A segunda manifesta-se na confiabilidade do treinamento sob regimes adversos, evidenciada pela bimodalidade do cenário de aperto monetário: sob condições desfavoráveis, uma fração substancial das inicializações não converge para política funcional, de sorte que a qualidade de uma execução individual não pode ser presumida.

O achado de maior valor, contudo, é a invariância da política entre regimes, observada ainda que as variáveis macroeconômicas compõem o espaço de estados. A hipótese mais parcimoniosa para o fenômeno não está em incapacidade intrínseca do algoritmo, mas na heterogeneidade de escala das observações fornecidas à rede. Como os atributos do estado não foram normalizados, alimentam a política em ordens de grandeza marcadamente distintas: enquanto os índices de risco-retorno situam-se em escala unitária, a taxa SELIC diária ocupa a ordem de 10^{-4} . Em uma entrada não normalizada, tal disparidade faz com que a variação do sinal de juros seja numericamente ofuscada pelas demais variáveis, tornando praticamente inacessível à política justamente a informação que deveria fundamentar a adaptação ao regime. Privado de sinal macroeconômico efetivo, o agente converge para uma política ativa, porém insensível ao contexto, e a condicionalidade documentada do desempenho decorre da interação dessa política estável com as condições de cada mercado. O impacto da normalização das observações sobre algoritmos de gradiente de política é amplamente documentado [Huang et al. 2022], o que confere ao diagnóstico plausibilidade mecânica e o converte na principal direção de trabalho futuro.

Estabelecido que a vantagem no cenário principal não decorre de adaptação macroeconômica, resta identificar sua origem. A composição média sobrepondera a renda variável (25,2%) e mantém posição levemente subponderada em ouro (19,3%), ao passo que, no período, o BOVA11 foi o ativo de menor retorno e o GOLD11 o de maior por larga margem. O tilt estratégico médio foi, portanto, desfavorável, e a superação observada não pode ser atribuída a posicionamento estratégico vantajoso: sua origem reside necessariamente na realocação de curto prazo, materializada no intenso comportamento transacional. O agente não atua como estrategista que define composição-alvo favorável, mas como

operador tático que extrai retorno do reposicionamento de alta frequência, capacidade distinta da adaptação macroeconômica e dela independente.

6. Conclusão

Este trabalho avaliou, sob protocolo multi-seed e em três regimes macroeconômicos, um agente PPO para o rebalanceamento dinâmico de uma carteira de ETFs brasileiros em ambiente que integra contexto macroeconômico e fricção transacional. A síntese das evidências sustenta conclusão dupla. No plano do desempenho, o método demonstrou viabilidade condicional ao regime, superando a estratégia passiva sob juros estáveis, cedendo-lhe sob choque agudo e sendo dominado pela renda fixa sob aperto monetário. No plano da política, a análise da composição revelou que essa condicionalidade não decorre de adaptação macroeconômica, e sim da interação de uma política aproximadamente invariante com as condições de cada mercado, caracterizando o agente como alocador tático direcional, e não como gestor defensivo ou estrategista macroeconômico.

As limitações documentadas definem o programa de pesquisa subsequente. A normalização das observações, ao tornar o sinal macroeconômico efetivamente acessível à política, permitiria reavaliar de forma justa a hipótese adaptativa; a adoção de recompensas que penalizem explicitamente o risco de queda mitigaria a vulnerabilidade observada na eficiência risco-retorno; e o desenvolvimento de mecanismos de detecção precoce de degeneração de sementes endereçaria a fragilidade do método sob transição abrupta. Em última análise, demonstra-se que a aplicação de Aprendizado por Reforço ao rebalanceamento de carteiras no mercado brasileiro é viável e competitiva sob condições específicas, mas está longe de constituir solução universal, e a compreensão madura de um sistema concreto, de suas capacidades e de seus limites reais, constitui fundamento mais sólido para o avanço da área do que a afirmação de uma superioridade que os dados não sustentariam.

Referências

- Almahdi, S. and Yang, S. Y. (2017). An adaptive portfolio trading system: a risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications*, 87:267–279.
- Almgren, R. and Chriss, N. (2000). Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40.
- Ang, A. (2014). *Asset Management: A Systematic Approach to Factor Investing*. Oxford University Press, Oxford.
- B3 — Brasil, Bolsa, Balcão (2024). Tarifas de ações e fundos de investimento. Disponível em: https://www.b3.com.br/pt_br/produtos-e-servicos/tarifas/.
- Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press, Princeton.

- Clinton, J. D. (2011). Proxy variable. In *The SAGE Encyclopedia of Social Science Research Methods*. SAGE Publications, Thousand Oaks.
- DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: how inefficient is the 1/N portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953.
- Gunjan, A. and Bhattacharyya, S. (2023). A brief review of portfolio optimization techniques. *Artificial Intelligence Review*, 56:3847–3886.
- Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. (2018). Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Huang, S., Dossa, R. F. J., Raffin, A., Kanervisto, A., and Wang, W. (2022). The 37 implementation details of proximal policy optimization. *The ICLR Blog Track*.
- Jiang, Z., Xu, D., and Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*.
- Leland, H. E. (1996). Beyond mean-variance: performance measurement in a nonsymmetric world. *The Journal of Finance*, 54(1):27–36.
- Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.
- Masters, S. J. (2003). Rebalancing. *The Journal of Portfolio Management*, 29(3):52–57.
- Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.
- Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-Baselines3: reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 2nd edition.
- Uppaladinni, S. et al. (2024). Machine learning for real-time portfolio rebalancing: a novel approach to financial optimization. *International Journal for Research in Applied Science & Engineering Technology*, 12(10).
- Vanguard Research (2019). *Best Practices for Portfolio Rebalancing*. Vanguard White Paper, Valley Forge.
- Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach*. Cengage Learning, Boston, 6th edition.