CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

APRENDIZADO POR REFORÇO PROFUNDO APLICADO À NEGOCIAÇÃO INTRADIÁRIA DE MÚLTIPLAS AÇÕES

Fernando Henrique de Jesus Fraga da Silva

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Eduardo Bezerra da Silva, D.Sc.
Pedro Henrique González Silva, D.Sc.

Rio de Janeiro,

Janeiro de 2026

Aprendizado por Reforço Profundo Aplicado à Negociação Intradiária de Múltiplas Ações

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Fernando Henrique de Jesus Fraga da Silva

Aprovada por:

Presidente, Eduardo Bezerra da Silva, D.Sc. (CEFET/RJ)

Pedro Henrique González Silva, D.Sc. (CEFET/RJ & COPPE/PESC)

Glauco Fiorott Amorim, D.Sc. (CEFET/RJ)

Aline Marins Paes Carvalho D.Sc (IC/UFF)

Rio de Janeiro,

Janeiro de 2026

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

S586 Silva, Fernando Henrique de Jesus Fraga da
Aprendizado por reforço profundo aplicado à negociação intradiária de múltiplas ações / Fernando Henrique de Jesus Fraga da Silva. — 2026.
86f. + apêndice : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2026.
Bibliografia : f. 83-86
Orientador: Eduardo Bezerra da Silva
Coorientador: Pedro Henrique González Silva

1. Aprendizado por reforço. 2. Aprendizado do computador. 3. Algorítmos computacionais. 4. Mercado financeiro. 5. Mercado de ações – Previsão. I. Silva, Eduardo Bezerra da. (Orient.). II. Silva, Pedro Henrique González. (Coorient.). III. Título.

CDD 006.3

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

DEDICATÓRIA

Dedico este trabalho aos meus pais, Edielcio e Rosemere, por todo o apoio, amor e ensinamentos ao longo da minha vida, que foram fundamentais para a formação de quem sou hoje. Dedico também ao amor da minha vida, Nathália, pelo incentivo constante, compreensão e apoio incondicional sempre.

AGRADECIMENTOS

Agradeço, primeiramente, aos meus pais, pelo suporte contínuo, pelos valores transmitidos e por sempre acreditarem em mim, mesmo nos momentos mais desafiadores.

À minha namorada, pelo companheirismo, paciência e apoio emocional ao longo desta jornada acadêmica, tornando os desafios mais leves e as conquistas mais significativas.

Aos meus orientadores, pela confiança, disponibilidade, orientação e valiosas contribuições que foram essenciais para o desenvolvimento e conclusão deste trabalho.

Aos meus companheiros de trabalho, pelo apoio, compreensão e incentivo durante o período do mestrado.

Aos amigos do PPCIC, pelas trocas de conhecimento, discussões enriquecedoras e pelo companheirismo ao longo dessa caminhada acadêmica.

Por fim, a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho, o meu sincero agradecimento.

RESUMO

Aprendizado por Reforço Profundo Aplicado à Negociação Intradiária de Múltiplas Ações

Fernando Henrique de Jesus Fraga da Silva

Orientadores:
Eduardo Bezerra da Silva, D.Sc.
Pedro Henrique González Silva, D.Sc.

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

O mercado de ações é um ambiente dinâmico e volátil, no qual empresas de capital aberto negociam frações de seu valor, sujeitas a variações contínuas de preço influenciadas por fatores econômicos, políticos e sociais. Antecipar essas variações é uma tarefa complexa, especialmente no contexto da negociação intradiária, em que as decisões de compra e venda precisam ser tomadas em intervalos de tempo muito curtos, com base em dados que mudam rapidamente. Nesse cenário, o Aprendizado por Reforço (AR) surge como um paradigma promissor, capaz de desenvolver estratégias adaptativas por meio da interação contínua entre agente e ambiente. Esta dissertação investiga o uso de técnicas de Aprendizado por Reforço Profundo (ARP) na negociação financeira, com foco em cenários de negociação intradiária envolvendo múltiplas ações. É proposta uma abordagem baseada em ARP para estimar ações de compra e venda simultaneamente em diversos ativos, utilizando dados de mercado de alta granularidade, de modo a aproximar-se das condições reais de negociação. As análises experimentais foram conduzidas com o algoritmo Proximal Policy Optimization (PPO). Os resultados obtidos indicam que o agente proposto apresentou desempenho superior em relação a estratégias tradicionais de referência, alcançando ganhos superiores a 10 pontos percentuais em determinados casos.

Palavras-chave:
Aprendizado por Reforço, Aprendizado por Reforço Profundo, PPO, Mercado Financeiro, Mercado de Ações, Negociação Intradiária

Rio de Janeiro,

Janeiro de 2026

ABSTRACT

Aprendizado por Reforço Profundo Aplicado à Negociação Intradiária de Múltiplas Ações

Fernando Henrique de Jesus Fraga da Silva

Advisors:
Eduardo Bezerra da Silva, D.Sc.
Pedro Henrique González Silva, D.Sc.

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

The stock market is a dynamic and volatile environment in which publicly traded companies negotiate fractions of their value, subject to continuous price fluctuations influenced by economic, political, and social factors. Anticipating these fluctuations is a complex task, especially in the context of intraday trading, where buy and sell decisions must be made within very short time intervals based on rapidly changing data. In this scenario, Reinforcement Learning (RL) emerges as a promising paradigm capable of developing adaptive strategies through the continuous interaction between agent and environment. This dissertation investigates the use of Deep Reinforcement Learning (DRL) techniques in financial trading, focusing on intraday scenarios involving multiple stocks. It proposes a DRL-based approach to estimate buy and sell actions simultaneously across various assets, using high-granularity market data to better approximate real trading conditions. Experimental analyses were conducted using the Proximal Policy Optimization (PPO) algorithm. The results indicate that the proposed agent outperformed traditional benchmark strategies, achieving gains exceeding 10 percentage points in certain cases.

Key-words:
Machine learning, Reinforcement learning, PPO, Financial market, Stock market, Day trading

Rio de Janeiro,

Janeiro de 2026

Sumário

Declaração de Uso de Ferramentas de IA 13

I Introdução 14

I.1 Contextualização 14

I.2 Objetivos 16

I.3 Justificativa 17

I.4 Metodologia 18

I.5 Organização dos Capítulos 18

II Referencial Teórico 20

II.1 Mercado Financeiro 20

II.1.1 Negociação intradiária 21

II.2 Aprendizado por reforço 21

II.2.1 Processo de Decisão de Markov 22

II.2.2 Conceitos básicos do Aprendizado por Reforço 22

II.2.2.1 *Exploration vs Exploitation* 23

II.2.2.2 Equações importantes 24

II.2.3 Principais Abordagens de Solução 25

II.2.4 Métodos Ator-Crítico 25

II.2.5 Teorema do Gradiente de Política 26

II.2.6 Métodos de Região de Confiança 27

II.2.7 Função Objetivo Substituta 28

II.2.8 Coeficiente de Penalidade KL Adaptável 29

II.2.9 Proximal Policy Optimization (PPO) 30

III Trabalhos Relacionados 32

III.1 Seleção dos Trabalhos Relacionados 32

III.1.1 Critérios de Inclusão 33

III.1.2 Critérios de Exclusão 33

III.1.3 Critérios de Priorização 33
III.1.4 Resultados 34
III.2 Síntese Comparativa 34

IV Conjuntos de Dados 41
IV.1 Obtenção dos dados 41
IV.2 Tratamento de dados 42
IV.3 Construção das Entradas do Agente 44
IV.3.1 Seleção das variáveis observadas 44
IV.3.2 Normalização e padronização 45
IV.3.3 Agregação temporal 45
IV.3.4 Representação final do estado 46
IV.4 Análise exploratória 46
IV.4.1 Análise de Retorno Percentual Acumulado 47
IV.4.1.1 Empresas de um mesmo setor 48
IV.4.1.2 Valorização de portfólio 50
IV.4.2 Análise de tendência 51

V Metodologia para Construção do Agente de Negociação Intradiária via Aprendizado por Reforço Profundo 55
V.1 Formulação do problema 55
V.2 Especificação do Processo de Decisão de Markov 57
V.2.1 Estados 57
V.2.2 Ações 58
V.2.3 Função Recompensa 59
V.2.3.1 Função Local 59
V.2.3.2 Função de Lucro Líquido 59
V.2.3.3 Função Diferencial de Lucro Líquido 59
V.2.4 Parâmetros da Modelagem 60
V.3 Treinamento e validação do agente 60
V.3.1 Configuração dos Hiperparâmetros do algoritmo 61
V.3.2 Validação Cruzada 62
V.3.3 Métricas utilizadas 63

VI Avaliação Experimental 66
VI.1 Configuração de Hardware e Software 66

VI.2 Experimentos de sensibilidade ao fator de desconto 67
VI.3 Experimentos de sensibilidade às variáveis do estado 69
VI.4 Comparação de arquiteturas de Aprendizado por Reforço 71
VI.5 Resultados 72
VI.5.1 Treinamento e teste para uma mesma ação 73
VI.5.2 Treinamento e teste para ações diferentes 78
VI.5.3 Treinamento e teste com um portfólio de ações 79

VII Conclusões **81**
VII.1 Análise retrospectiva 81
VII.2 Trabalhos futuros 82
Referências 83

A Sumários Estatísticos dos Conjuntos de Dados Coletados **87**

Lista de Figuras

II.1 Diagrama do ciclo de Aprendizado por Reforço 23

II.2 Representação as trajetórias do gradiente de política 27

III.1 Fluxograma de trabalhos relacionados 34

IV.1 Arquitetura de coleta de dados do *Chicago Board Options Exchange* (CBOE) 42

IV.2 *Layout* do site CBOE 43

IV.3 Valorização das empresas de tecnologia 49

IV.4 Valorização de instituições financeiras 50

IV.5 Valorização de empresas de linhas aéreas 51

IV.6 Valorização de empresas automotivas 52

IV.7 Valorização de empresas de mineração 53

IV.8 Valorização de empresas de diversos setores 53

IV.9 Valorização do portfólio com as 19 empresas 54

V.1 Processo de decisão do algoritmo 56

V.2 Validação cruzada para séries temporais 62

VI.1 Preço e lucro por transição para ação da Coca-Cola 74

Lista de Tabelas

III.1 Palavras-chave para encontrar artigos relacionados 32

III.2 Resumo de trabalhos em aprendizado por reforço 36

III.3 Resumo de trabalhos em aprendizado por reforço 37

IV.1 Dados coletados da empresa Apple 43

IV.2 Dados coletados da empresa Tesla 44

IV.3 Variáveis utilizadas na construção do estado do agente 47

IV.4 Empresas analisadas com respectivos ramos e sumário estatístico 48

IV.5 Tendência das ações 54

V.1 Parâmetros do modelo utilizados na definição do Processo de Decisão de Markov (PDM) 60

V.2 Divisão das dobras na validação cruzada temporal 63

VI.1 Tabela de softwares utilizados no scrapper 67

VI.2 Tabela de bibliotecas e frameworks utilizados no treinamento e teste do agente 68

VI.3 Experimentos de sensibilidade ao fator de desconto 68

VI.4 Experimentos de sensibilidade às variáveis do estado das informações 70

VI.5 Teste de comparação de arquiteturas de Aprendizado por Reforço 72

VI.6 Resultados diários treinando e testando o agente para uma mesma ação 75

VI.7 Resultados dos testes estatísticos por empresa 76

VI.8 Resultados do agente para treinamento em uma ação e teste em outra 78

VI.9 Resultados do agente sendo treinado e testado em um portfólio 80

A.1 Estatísticas descritivas das ações de empresas de tecnologia 87

A.2 Estatísticas descritivas das ações de empresas do setor financeiro 88

A.3 Estatísticas descritivas das ações de empresas do setor aéreo 89

A.4 Estatísticas descritivas das ações de empresas do setor automotivo 90

A.5 Estatísticas descritivas das ações de empresas do setor de mineração 91

A.6 Estatísticas descritivas das ações de empresas de setores diversos 92

Lista de Abreviações

A2C	*Advantage Actor-Critic*	58, 60, 61, 71, 81
A3C	*Asynchronous Advantage Actor-Critic*	25
AM	Aprendizado De Máquina	15, 16, 17, 21, 33, 69
AP	Aprendizado Profundo	33
AR	Aprendizado Por Reforço	16, 17, 20, 21, 22, 23, 33, 55, 68, 73, 77, 78, 80
ARP	Aprendizado Por Reforço Profundo	16, 17, 34, 41, 55, 70, 81, 82
CBOE	*Chicago Board Options Exchange*	17, 18, 41, 42, 43, 44, 81
CNN	*Convolutional Neural Network*	37
DDPG	*Deep Deterministic Policy Gradient*	36, 37, 38, 58, 60, 61, 71, 81
DPG	*Deterministic Policy Gradient*	71
DQN	*Deep Q Learning*	36, 37, 71
IA	Inteligência Artificial	14
KL	*Kullback-Leibler*	27, 28, 29, 30
LSTM	*Long Short-Term Memory*	37
NN	*Neural Network*	36
PDM	Processo De Decisão De Markov	16, 17, 20, 22, 26, 56, 57, 60, 61
PPO	*Proximal Policy Optimization*	25, 29, 37, 38, 58, 60, 61, 71, 72, 81, 82
SAC	*Soft Actor-Critic*	58, 60, 61, 71, 72, 81
SGBD	Sistema Gerenciador De Banco De Dados	42
TRPO	*Trust Region Policy Optimization*	27, 29

Declaração de Uso de Ferramentas de Inteligência Artificial

O autor declara que utilizou ferramentas de Inteligência Artificial generativa exclusivamente como apoio à produção textual desta dissertação, incluindo sugestões de redação, revisão linguística, esclarecimento de conceitos técnicos e organização de ideias. Todas as decisões conceituais, análises, interpretações, escolhas metodológicas, resultados experimentais e conclusões apresentadas são de responsabilidade integral do autor.

Nenhum conteúdo gerado por ferramentas de IA foi incorporado sem revisão crítica. O uso dessas ferramentas ocorreu em conformidade com princípios de integridade acadêmica, transparência e responsabilidade, e não substituiu o trabalho intelectual necessário à elaboração desta pesquisa.

Capítulo I Introdução

I.1 Contextualização

O mercado de ações é uma parte essencial de qualquer economia. Esse mercado constitui o ambiente no qual as empresas de capital aberto negociam frações de seu capital. O mercado de ações é considerado um mercado de alto risco, pois qualquer mudança política, econômica, movimentos de outros mercados financeiros ao redor do mundo, catástrofes naturais, guerras, entre outros diversos fatores, podem provocar reações nele, que podem afetar tanto a economia local quanto mundial [Altuner and Kilimci, 2022].

Com a evolução computacional ao longo dos anos, o mercado de ações mostra ter um grande potencial para utilizar modelos de Inteligência Artificial (IA) em previsões mais precisas, por duas principais razões. Primeiro, a eficácia dos modelos de IA para esse tipo de mercado já é objeto de evidência comprovada, conforme investigações recentes [Conegundes and Pereira, 2020; Zhang et al., 2023; Qiu et al., 2021]. Segundo, a IA consegue processar vastas quantidades de dados, o que aprimora suas capacidades preditivas [Gülmez, 2023; Milana and Ashta, 2021].

O mercado financeiro é um dos principais setores onde a IA pode ser aplicada. Isso ocorre porque, além da precisão dos modelos, é crucial lidar com uma enorme quantidade de dados a serem processados. Milhões de investidores globalmente almejam maximizar seus lucros, independentemente do ativo negociado ou da praça financeira utilizada. Nesse contexto, previsões precisas são essenciais para decisões de investimento informadas. A análise preditiva pode auxiliar em questões-chave, como quais ações escolher, quando e quanto investir. Uma maior taxa de acerto nas previsões pode resultar em um retorno sobre o investimento (ROI) mais elevado, sendo um fator decisivo para investidores que buscam maximizar seus lucros [Hu et al., 2021; Théate and Ernst, 2021].

Apesar de ser um tema amplamente estudado ao longo das últimas décadas [Iyer and Mehra, 2018; Heaton et al., 2016; Atsalakis and Valavanis, 2009; Kumar et al., 2021], prever com precisão os preços futuros das ações e criar estratégias de negociação lucrativas com base nessas previsões ainda representa um desafio considerável [Conegundes and Pereira, 2020]. A previsão é uma das atividades mais desafiadoras e quando relacionada ao mercado financeiro torna-se ainda mais

complexa devido aos riscos adicionais envolvidos [Chou and Truong, 2019]. A dificuldade em prever os valores futuros dessas séries decorre de seu caráter dinâmico e não linear, influenciado por variáveis que estão em constante mudança [Coelho et al., 2008].

Os preços de ações e moedas podem ser influenciados por uma ampla gama de fatores, como notícias, economia, política e outro fatores que são frequentemente complexos e difíceis de analisar [Maneejuk and Srichaikul, 2021; Wang et al., 2022b]. Outro fator são que os preços dos mercados financeiros estão em constante mudança, e as flutuações podem ocorrer a qualquer momento. Assim, prever os movimentos do mercado exige uma compreensão sutil desses fatores complexos [Wang et al., 2022a]. Em terceiro lugar, as negociações podem ser influenciadas por sentimentos como medo e ganância [Lin et al., 2018]. Além disso, a alta volatilidade e a natureza ruidosa do mercado financeiro torna extremamente difícil prever os preços das ações com precisão [Sun et al., 2023]. Nesse ambiente complexo, o objetivo consiste em maximizar o lucro da venda de ações ou vendê-las antes de o valor ter uma queda grande [Wu et al., 2020].

Técnicas provenientes de diferentes áreas têm sido amplamente utilizadas em aplicações no mercado financeiro, incluindo Otimização, Processamento de Sinais, Análise de Séries Temporais, Teoria de Controle, Métodos Estatísticos Avançados e Aprendizado de Máquina (AM), com interesse especial e crescente da comunidade financeira nesta última [Conegundes and Pereira, 2020]. Nesse sentido, a utilização de técnicas de AM se mostram validas, pois as decisões de compra e venda no mercado financeiro podem ser auxiliadas pelas previsões feitas pelos algoritmos de AM. O uso de máquinas para tomar essas decisões tem aumentado constantemente no mercado financeiro [Khushi and Meng, 2019].

O uso mais tradicional de AM para negociar no mercado de ações é aplicar técnicas de aprendizado supervisionado para prever preços futuros em um problema de regressão. Essas previsões são então usadas por um sistema de negociação que compra e vende ações com base em regras predefinidas para operar entradas e saídas de posições [Conegundes and Pereira, 2020; Khushi and Meng, 2019].

Segundo Conegundes and Pereira [2020], os principais problemas dessa abordagem são:

1. ela depende fortemente da qualidade das previsões, mas prever preços futuros é muito difícil;

2. as previsões de preços não se traduzem diretamente em sinais de compra e venda, exigindo uma camada adicional de lógica para convertê-las em ações de mercado;

3. só temos um sistema de negociação totalmente automatizado se tivermos pelo menos dois subsistemas, onde a saída do primeiro (o modelo de aprendizado supervisionado) é a entrada do segundo (o sistema de negociação);

4. o propósito do modelo de aprendizado (minimizar o erro de previsão) é diferente do propósito do sistema de negociação completo, que é maximizar o lucro; e

5. como as regras de negociação não são incorporadas na definição do algoritmo de aprendizado, várias restrições importantes do mundo real são desconsideradas nesse primeiro subsistema "inteligente".

Quando consideramos métodos de AM específicos, os métodos Aprendizado por Reforço (AR) se mostram eficientes, pois seu objetivo é maximizar alguma parte da recompensa cumulativa [Wu et al., 2020]. O aprendizado por reforço visa o aprendizado por meio de tentativa e erro [Sun et al., 2023], usando um *feedback* de suas experiências e otimizando a próxima iteração [Altuner and Kilimci, 2022]. Com uso dessa técnica, podemos treinar algoritmos por meio da otimização das funções de recompensa específicas para cada tarefa [Sun et al., 2023].

Segundo Sun et al. [2023], há quatro vantagens para utilizar aprendizado por reforço para o mercado de ações:

1. aprendizado por reforço permite treinar um agente de ponta a ponta, recebendo informações do mercado como entrada e ações de negociação de saída;

2. aprendizado por reforço ignora a tarefa de prever o preço futuro e otimizar o lucro geral diretamente;

3. restrições específicas podem ser facilmente utilizadas em aprendizado por reforço;

4. os métodos de aprendizado por reforço têm potencial para generalizar para qualquer condição de mercado.

Neste trabalho, um algoritmo de Aprendizado por Reforço Profundo é usado para otimizar carteiras com diversas ações. Para isso, um ambiente de transações é modelado como um PDM. Vários fluxos de valores financeiros de entrada são simulados.

I.2 Objetivos

A pergunta de pesquisa que investigamos neste trabalho é a seguinte: ***Algoritmos de Aprendizado por Reforço Profundo (ARP) são adequados para problemas de negociação intradiária de múltiplas ações?***

O objetivo principal deste trabalho é propor uma técnica de Aprendizado por Reforço Rrofundo (ARP) para estimar transações de negociação intradiária para diversas empresas, com uma

granularidade de transações o mais próximo da realidade do mercado possível. Para isso, o estudo concentra-se no treinamento de um agente de ARP com dados de negociações provenientes do site da CBOE, uma empresa especializada em ações e soluções de *trading* e investimentos.

Como objetivos secundários, podemos listar:

- Modelar corretamente as transações de compra e venda de ações como um PDM, de modo que possa ser resolvido com um algoritmo de aprendizado por reforço profundo;
- Comparar o desempenho do algoritmo escolhido com métodos utilizados em outros estudos;
- Conduzir uma análise experimental (por meio de simulações) do comportamento do algoritmo para diferentes janelas temporais.

As contribuições esperadas por este trabalho são:

- Um algoritmo de aprendizado por reforço profundo que seja capaz de fazer previsões de compra e venda de multi-ações em modelo de negociação intradiária, com resolução temporal de um minuto;
- Uma arquitetura para captura de dados do CBOE, com persistência de dados no formato Parquet e em banco de dados.

I.3 Justificativa

O uso de AR para resolver o problema proposto nesse trabalho é justificado por algumas razões [Conegundes and Pereira, 2020]:

1. otimização de ponta a ponta. Diferente do método de AM supervisionado, onde há diversas etapas para serem feitas, como análise de dados, treinamento de modelos supervisionados, otimização de parâmetros, etc., no AR essas etapas são feitas pelo próprio algoritmo;
2. criação automática e inteligente do sistema de negociação. o AR aprende diretamente uma política (sistema de negociação), sendo capaz de aprender políticas mais complexas do que aquelas derivadas manualmente por especialistas; e
3. treinamento diretamente em ambientes de simulação. o treinamento do modelo já é no ambiente de simulação, considerando fatores tais como liquidez, latência, *slippage* e custos de transação.

Outra motivação deste estudo é que trabalhos que utilizam métodos de aprendizado por reforço em mercado financeiro normalmente utilizam uma representação dos dados de forma resumida, considerando somente resumos dos valores ou quantidades negociadas (como valores de

abertura, fechamento, volume, cotação máxima, cotação mínima, etc.). Este trabalho propõe um método que possui uma granularidade temporal elevada, sendo esta granularidade de dados coletados de transações a cada minuto por companhias de diversos setores da economia americana.

I.4 Metodologia

A fim de criar e treinar um algoritmo capaz de realizar decisões intra-diárias de compra e venda de ações, um conjunto de etapas são necessárias.

A etapa inicial consistiu na definição do Processo de Decisão de Markov (MDP) a ser empregado. Nesse momento, foram estabelecidas todas as condições iniciais do problema, bem como especificadas as funções de recompensa, as políticas adotadas, os critérios de término de episódio e os espaços de observação e de ação atribuídos ao agente. As particularidades da versão atual dessa configuração são detalhadas no Capítulo V.

A etapa subsequente consiste na coleta dos dados, a partir dos dados disponibilizados no CBOE. Para tal finalidade, foi desenvolvido um sistema de extração automatizada utilizando a linguagem de programação Python, o qual extrai informações do site de forma contínua em um horário pré-estabelecido. Com essa obtenção há uma etapa de tratamento e preparação dos dados. Este processo envolveu a padronização de tipos numéricos, ajustes em formatações decimais, e a remoção de registros incompletos ou incorretos. Adicionalmente, foram aplicados filtros aos dados para a eliminação de duplicidades, manutenção apenas das colunas relevantes e exclusão de informações que não contribuíam para o treinamento e avaliação do algoritmo.

A etapa subsequente consiste no treinamento do agente. Com os dados já coletados e devidamente tratados, o modelo é alimentado por essas informações, o que possibilita a análise dos estados, ações, recompensas associadas às decisões tomadas pelo agente e os critérios de finalização dos episódios. Esses elementos possibilitam a interação adequada do agente com o ambiente, orientando-o na adoção de estratégias mais eficazes.

Como estratégia para avaliar o desempenho do agente, serão utilizados benchmarks do mercado para entender se o retorno do agente é superior aos índices utilizados no mercado. Para garantir a validade dos testes, serão utilizadas observações distintas daquelas usadas na fase de treinamento, evitando viés nos resultados.

I.5 Organização dos Capítulos

Além desta introdução, este trabalho possui mais cinco capítulos. O Capítulo II apresenta o referencial teórico. No Capítulo III apresentamos os trabalhos relacionados. No Capítulo V

descrevemos detalhes da metodologia para treinamento dos agentes de negociação intradiária. No Capítulo IV, descrevemos o conjunto de dados empregado neste trabalho, apresentando a análise exploratória e as transformações realizadas. No Capítulo VI fazemos a avaliação do modelo proposto. Finalmente, o Capítulo VII apresentamos a conclusão, bem como algumas ideias para trabalhos futuros. Complementando o conteúdo principal, este trabalho também inclui o Apêndice A, onde são apresentados os sumários estatísticos para cada um dos conjuntos de dados, contendo dados estatísticos para cada uma das colunas utilizadas

Capítulo II Referencial Teórico

Este capítulo apresenta alguns fundamentos necessários para o entendimento dos capítulos seguintes. A Seção II.1 introduz o funcionamento do mercado financeiro, destacando seus principais segmentos, mecanismos de precificação, desafios enfrentados e discute a prática de negociação intradiária. A Seção II.2 aborda os principais conceitos de Aprendizado por Reforço (AR), incluindo a formulação por meio de Processo de Decisão de Markov (PDM), o ciclo de interação agente-ambiente, o equilíbrio entre *exploration* e *exploitation*, bem como as funções de valor e vantagem.

II.1 Mercado Financeiro

O mercado financeiro é um sistema complexo e dinâmico onde ocorrem transações de ativos financeiros, como ações, títulos de dívida pública e privada, commodities, fundos de investimento, câmbio, entre outros instrumentos financeiros [Polleit, 2021]. Essas negociações se realizam em ambientes organizados, como bolsas de valores, ou no mercado de balcão, sendo fundamentais para a alocação eficiente de recursos na economia e para a precificação de ativos baseada nas expectativas dos agentes econômicos [Altuner and Kilimci, 2022].

Esse ambiente é segmentado em diferentes mercados, como o mercado de capitais, onde são negociadas ações e debêntures; o mercado monetário, que lida com ativos de curtíssimo prazo e elevada liquidez; o mercado de crédito, que envolve operações de empréstimos e financiamentos; e o mercado de câmbio, no qual são feitas operações de compra e venda de moedas estrangeiras. Cada segmento possui seus próprios mecanismos de operação, participantes, riscos e regulamentações [Hull, 2014; Wang et al., 2022b; Lin et al., 2018].

A formação dos preços dos ativos nesse mercado está diretamente relacionada à oferta e demanda, sendo influenciada por variáveis macroeconômicas, geopolíticas, comportamento dos investidores, resultados corporativos e indicadores técnicos [Altuner and Kilimci, 2022]. Devido a essa complexidade, investidores e instituições utilizam modelos quantitativos e ferramentas computacionais para embasar suas decisões [Wilmott, 2007; Gülmez, 2023; Milana and Ashta, 2021].

Atualmente, um dos principais desafios enfrentados por participantes do mercado financeiro,

como investidores institucionais e individuais, é a otimização da relação risco-retorno. De forma mais precisa, busca-se estruturar estratégias que maximizem o retorno esperado de um portfólio ou operação, ao mesmo tempo, em que se controla ou minimiza o nível de risco assumido [Santos and Coelho, 2010]. Entre as estratégias empregadas com esse propósito, destaca-se a negociação intradiária, que visa aproveitar variações intradiárias de preço para obter retorno financeiro [Kyröläinen, 2008].

II.1.1 Negociação intradiária

A negociação intradiária é uma modalidade de operação especulativa na qual o investidor abre e encerra posições no mesmo ativo dentro do mesmo pregão, ou seja, no intervalo de um único dia de negociação [Hull, 2014]. O objetivo é capturar movimentos de curto prazo nos preços dos ativos, como ações ou contratos futuros, explorando padrões técnicos, indicadores, volume ou outros sinais de mercado. Essa prática exige alta frequência de decisões, capacidade de resposta rápida às flutuações do mercado e, em muitos casos, o uso de sistemas automatizados de negociação [Kyröläinen, 2008].

Apesar do potencial de altos retornos, a negociação intradiária também está associado a elevados riscos, especialmente devido à sua alta alavancagem, elevada frequência de transações e sensibilidade a ruídos de mercado [Kyröläinen, 2008; Conegundes and Pereira, 2020].

Devido à elevada volatilidade e à complexidade envolvida na previsão de preços em horizontes temporais tão curtos, o uso de técnicas computacionais avançadas tornou-se cada vez mais recorrente. Em especial, técnicas de aprendizado de máquina têm sido empregadas como ferramentas auxiliares no processo decisório, permitindo a modelagem de comportamentos de mercado com base em dados históricos e em tempo real. Dentre essas abordagens, destaca-se o AR, um paradigma do AM que modela a interação de um agente com o ambiente por meio de ações que maximizam recompensas acumuladas ao longo do tempo [Conegundes and Pereira, 2020].

II.2 Aprendizado por reforço

O aprendizado por reforço (AR) é uma técnica de aprendizado de máquina cujo objetivo é encontrar uma política ótima que maximize o retorno esperado, sendo esse retorno a soma descontada das recompensas [Sutton and Barto, 2018]. Nesse método, não é informado para o algoritmo quais ações ele deve tomar, mas, em vez disso, o mesmo deve descobrir quais ações rendem a maximização da função de recompensa [Sutton and Barto, 2018]. Este tipo de aprendizado é baseado na psicologia comportamental dos seres vivos, já que esses interagem com o ambiente, aprendendo a partir de experiências vividas. No caso do algoritmo de aprendizado

por reforço, chamamos as experiências positivas de recompensas e as experiências negativas de punições [Taghavi et al., 2023].

Para alcançar esse objetivo, o agente pode testar diferentes políticas durante a fase de treinamento, sendo que apenas uma dessas políticas é a que maximiza a recompensa, chamada de política ótima [Carta et al., 2021].

Diferente do aprendizado supervisionado e não supervisionado, no AR o agente aprende a maximizar uma recompensa em um ambiente, em vez de aprender a partir de anotações (como no aprendizado supervisionado) ou agrupar dados similares (como no aprendizado não supervisionado) [Carta et al., 2021].

O AR pode ser formalizado por meio de um PDM.

II.2.1 Processo de Decisão de Markov

Um PDM representa uma estrutura matemática utilizada para modelar problemas de decisão sequencial, nos quais as escolhas realizadas em um determinado momento afetam tanto as recompensas imediatas quanto os estados futuros do sistema [Sutton and Barto, 2018].

Esse modelo é composto por um conjunto de estados possíveis, representado por S, e por um conjunto de ações disponíveis, representado por A. Define-se um estado inicial $s_1 \in S$, determinado por uma função de densidade de probabilidade $p_1(s_1)$. Além disso, há uma função de recompensa $r : S \times A \times S \to \mathbb{R}$, responsável por atribuir um retorno obtido ao realizar uma ação em um estado e transitar para outro. Também é definida uma função de transição de estados $p(s_{t+1} \mid s_t, a_t)$, que descreve a probabilidade de transição do estado s_t para o estado s_{t+1} após a execução da ação a_t, respeitando a propriedade de Markov de primeira ordem, expressa por $p(s_{t+1} \mid s_1, a_1, \ldots, s_t, a_t) = p(s_{t+1} \mid s_t, a_t)$. Por fim, o processo apresenta uma condição de término que determina seu encerramento quando satisfeita.

Para tomar decisões, é utilizada uma função de política $\pi_\theta : S \to A$, que define a ação a ser tomada com base no estado observado. Essa política pode ser estocástica, quando modelada por uma distribuição de probabilidade $\pi_\theta(a_t \mid s_t)$, ou determinística, quando definida por uma função direta $a_t = \mu_\theta(s_t)$. Em ambos os casos, o vetor de parâmetros $\theta \in \mathbb{R}^n$ representa os pesos ou coeficientes que ajustam a política.

II.2.2 Conceitos básicos do Aprendizado por Reforço

No aprendizado por reforço, o agente, que representa a entidade responsável por tomar decisões, interage continuamente com um ambiente que simula o contexto do problema a ser resolvido. A cada instante de tempo t, o agente observa um estado s_t e seleciona uma ação a_t com base

em sua política corrente. Em resposta a essa ação, o ambiente fornece uma nova observação de estado s_{t+1} e um valor de recompensa r_{t+1}.

Esse ciclo de interação se repete até que uma condição terminal seja atingida, encerrando o episódio. A sequência completa de estados, ações e recompensas gerada ao longo dessas interações é denominada *trajetória* e é simbolizada por τ.

A Figura II.1 ilustra esse processo, mostrando o fluxo de informações entre o agente e o ambiente: a cada passo, o agente observa o estado (s_t), escolhe uma ação (a_t), recebe uma recompensa (r_t) e transita para um novo estado (s_{t+1}), construindo assim a trajetória ao longo do episódio.

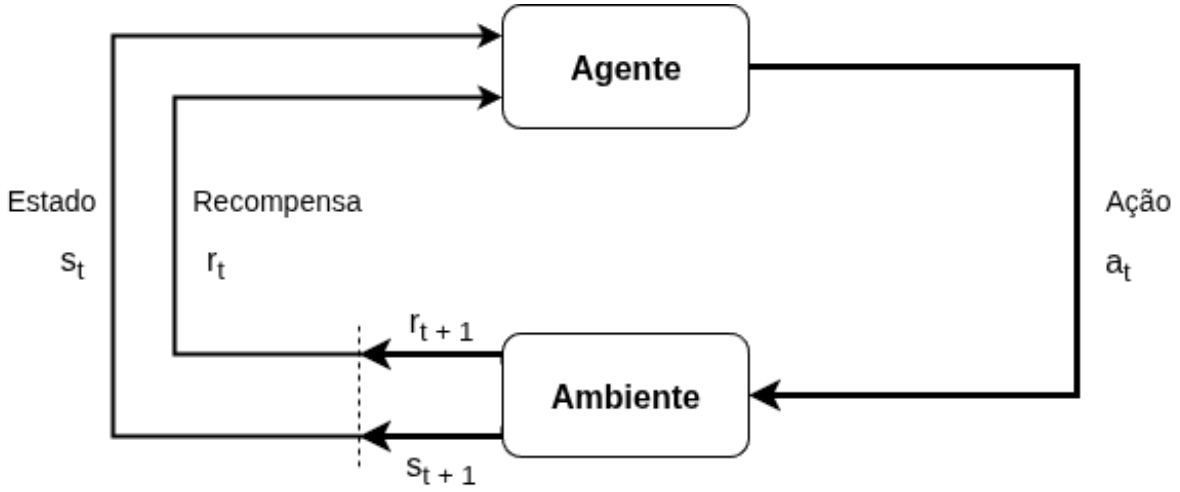


Figura II.1: Diagrama de blocos ilustrando a estrutura fundamental de um agente de AR. A figura mostra como, a cada passo temporal, o agente observa o estado do ambiente, seleciona uma ação e, em seguida, recebe uma recompensa e o novo estado resultante. Esse ciclo contínuo de interação compõe o processo de tomada de decisão que orienta o aprendizado do agente. Adaptado de [Sutton and Barto, 2018].

II.2.2.1 *Exploration vs Exploitation*

No contexto do aprendizado por reforço, existe um desafio central que consiste em equilibrar a exploração de novas ações (*Exploration*) com a exploração do conhecimento já adquirido pelo agente (*Exploitation*) [Sutton and Barto, 2018]. Se o agente sempre escolher a ação que, até o momento, apresentou o melhor retorno, ele poderá ignorar outras alternativas que, com o tempo, poderiam oferecer resultados superiores.

Por outro lado, se o agente apenas explorar novas ações, sem considerar as que já demonstraram bom desempenho, ele pode não alcançar uma solução eficiente, incorrendo em perda de recompensas maiores. Esse equilíbrio é um requisito para que o agente aprenda uma política eficaz.

II.2.2.2 Equações importantes

O retorno associado a uma trajetória τ, representado por $G(\tau)$, também é chamado de recompensa acumulada com desconto. Ele consiste na soma das recompensas recebidas ao longo do tempo, ponderadas por um fator de desconto $\gamma \in [0, 1]$. Esse fator é aplicado para reduzir a influência de recompensas futuras, refletindo a preferência por ganhos mais imediatos e evitando divergências em trajetórias infinitas [Sutton and Barto, 2018]. A fórmula correspondente está expressa na Equação II.1.

$$G(\tau) = \sum_{t=1}^{T} \gamma^t r_t \tag{II.1}$$

O propósito central dos algoritmos de aprendizado por reforço é encontrar uma política ótima que maximize a expectativa do retorno, $\mathbb{E}[G(\tau)]$.

Dentro desse contexto, destacam-se as funções de valor. A função de valor de uma política π em um estado s_t, denotada por $V^\pi(s_t)$, quantifica a expectativa do retorno acumulado a partir de s_t, assumindo que o agente seguirá a política π. Essa ideia é formalizada na Equação II.2.

$$V^\pi(s_t) = \mathbb{E}_\pi[G_t \mid s_t] \tag{II.2}$$

De modo semelhante, define-se a função valor-ação, $Q^\pi(s_t, a_t)$, que estima a recompensa esperada ao se tomar uma ação a_t no estado s_t, e então seguir a política π a partir do próximo estado. Essa função mede a utilidade de realizar determinada ação antes de seguir a política e está representada pela Equação II.3.

$$Q^\pi(s_t, a_t) = \mathbb{E}_\pi[G_t \mid s_t, a_t] \tag{II.3}$$

Finalmente, a função de vantagem indica quanto se pode esperar de retorno adicional caso o agente escolha uma ação arbitrária a_t em vez de seguir a política π no estado s_t. Ela é usada para avaliar como o retorno esperado seria alterado (aumentado ou diminuído) se a ação sugerida pela política π fosse substituída por uma ação diferente. A função de vantagem é calculada pela diferença entre a função de valor-ação e a função de valor. A Equação II.4 expressa o cálculo da função de vantagem.

$$A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t) \tag{II.4}$$

II.2.3 Principais Abordagens de Solução

O principal objetivo dos algoritmos de aprendizado por reforço é identificar a política que maximize a expectativa do retorno total descontado, a partir de um estado inicial s_0. Existem diferentes formas de alcançar essa meta, geralmente classificadas em três grupos principais:

1. Métodos Tabulares. São indicados para espaços de estados e ações discretos (ou passíveis de discretização). Nesses métodos, as funções de valor são representadas por tabelas e atualizadas iterativamente por meio das equações de Bellman, conforme Equações II.5 e II.6.

$$V^{\pi}(s_{t+1}) = r_t + \gamma V^{\pi}(s_t) \tag{II.5}$$

$$Q^{\pi}(s_{t+1}, a_{t+1}) = r_t + \gamma Q^{\pi}(s_t, a_t) \tag{II.6}$$

   A política ótima é inferida implicitamente: a melhor ação é aquela que maximiza V^{π} ou Q^{π} no estado atual. Um exemplo clássico é o algoritmo Q-Learning [Watkins and Dayan, 1992].

2. Otimização de Políticas. Nesta abordagem, a política é parametrizada diretamente, podendo ser estocástica, como $\pi_\theta(a_t|s_t)$, ou determinística, como $a_t = \mu_\theta(s_t)$. Essa abordagem é mais adequada a espaços contínuos de estados e ações. O algoritmo REINFORCE [Williams, 1992] é um exemplo representativo dessa categoria.

3. Algoritmos Ator-Crítico. Esta abordagem combina as ideias anteriores, sendo uma das mais utilizadas atualmente. Envolve dois componentes: o *ator*, que representa a política do agente, e o *crítico*, que estima a função valor-ação $Q^{\pi}(s, a)$ e orienta a atualização do ator. Um exemplo notável é o algoritmo *Asynchronous Advantage Actor-Critic* (A3C) [Mnih et al., 2016].

O algoritmo empregado neste trabalho, o *Proximal Policy Optimization* (PPO), pertence ao terceiro grupo.

II.2.4 Métodos Ator-Crítico

Os métodos ator-crítico combinam o uso do gradiente de política, apresentado na Seção II.2.5, com um segundo modelo aproximador para estimar a função $Q^{\pi}(s, a)$.

A estrutura pode ser interpretada como um sistema composto por um ator (o agente responsável por tomar decisões) e um crítico (uma função que avalia o valor das ações), que fornece feedback sobre a qualidade das escolhas feitas. A cada etapa do processo de aprendizado, tanto os parâmetros do ator quanto os do crítico são ajustados.

O ator pode operar de forma estocástica ou determinística. No modo estocástico, as ações são escolhidas com base em uma distribuição de probabilidade $\pi_\theta(a_t|s_t)$, conforme apresentado na Seção II.2.5. Já no modo determinístico, a ação é definida diretamente por uma função determinística: $a_t = \mu_\theta(s_t)$. Para empregar essa estratégia, é necessário compreender o teorema apresentado na Seção II.2.5

II.2.5 Teorema do Gradiente de Política

Assim como outros algoritmos de aprendizado por reforço, ele resolve problemas formulados como PDM, conforme descrito na Seção V.2.

Nesse contexto, o algoritmo opera considerando uma distribuição de probabilidade sobre o estado inicial, denotada por $p(s_0)$, e uma trajetória τ constituída por uma sequência ordenada de estados e ações, iniciando em s_0 e terminando em um estado terminal, de forma que $\tau = s_0, a_0, s_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T$. Cada transição ao longo dessa trajetória é associada a um valor real por meio de uma função de recompensa $R : S \times A \times S \to \mathbb{R}$. Além disso, o processo decisório é guiado por uma política estocástica $\pi_\theta(a_t|s_t)$, parametrizada por um vetor θ, a qual expressa a probabilidade de selecionar a ação a_t dado o estado s_t. Por fim, o comportamento dinâmico do ambiente é descrito pela função de transição de estados $p(s_{t+1}|s_t, a_t)$, que especifica a probabilidade de transitar do estado atual para o próximo estado após a execução da ação correspondente.

A probabilidade de uma trajetória completa τ pode ser expressa como o produto das probabilidades individuais de cada transição, conforme mostrado na Equação II.7 [Sutton and Barto, 2018].

$$p_\theta(\tau) = p(s_0) \cdot \pi_\theta(a_0|s_0) \cdot p(s_1|s_0, a_0) \cdots \pi_\theta(a_{T-1}|s_{T-1}) \cdot p(s_T|s_{T-1}, a_{T-1}). \tag{II.7}$$

Busca-se ajustar os parâmetros θ da política $\pi_\theta(a_t|s_t)$ de modo a maximizar a expectativa da recompensa acumulada e descontada ao longo das trajetórias, como expressa na Equação II.8.

$$\theta^* = \arg\max_\theta \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_t \gamma^t r_t(s_t, a_t) \right]. \tag{II.8}$$

A Figura II.2 ilustra a intuição do algoritmo: três trajetórias distintas partem do mesmo

estado inicial s_0, mas como são geradas de maneira estocástica, resultam em retornos diferentes. As regiões de maior intensidade, sinalizado pelas cores quentes, são os com recompensa mais alta, as regiões de menor intensidade, sinalizado pelas cores frias, são os com a recompensa mais baixa. A trajetória com o marcador verde representa o melhor retorno, seguida da em amarelo e da em vermelho. O objetivo do algoritmo é ajustar a política para aumentar a probabilidade de gerar trajetórias semelhantes àquela de maior retorno.

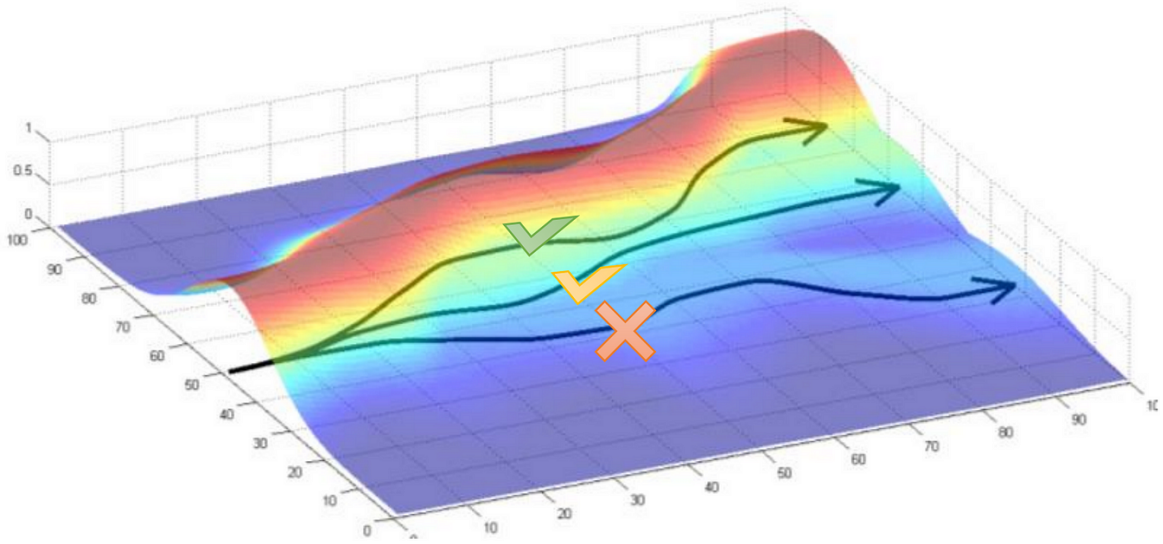


Figura II.2: A imagem apresenta uma representação esquemática da política de gradiente no contexto do aprendizado por reforço. O objetivo do agente é encontrar a trajetória ótima que maximize o acúmulo de recompensas ao longo do tempo. A trajetória sinalizada com um símbolo de confirmação verde indica uma política que conduz o agente em direção a regiões de maior recompensa. Já a trajetória com a marcação laranja representa uma política que, embora se aproxime de áreas recompensadoras, não é tão eficiente quanto a anterior. Por fim, a trajetória sinalizada com um "X" vermelho ilustra uma política subótima, que direciona o agente para regiões com menor recompensa. Fonte: Policy Gradients - CS 294-112: Deep Reinforcement Learning - Berkeley

II.2.6 Métodos de Região de Confiança

Os métodos de região de confiança foram introduzidos no contexto de aprendizado por reforço para aumentar a estabilidade das atualizações de política. Diferente dos métodos de gradiente de políticas tradicionais, que podem gerar atualizações agressivas e instáveis, os métodos de região de confiança impõem uma restrição explícita sobre a distância entre a política atual e a política anterior, promovendo uma melhoria mais conservadora e monotônica no desempenho.

Um dos algoritmos mais proeminentes que emprega essa abordagem é o *Trust Region Policy Optimization* (TRPO), proposto por [Schulman et al., 2017]. Esse método maximiza uma função objetivo substituta sob a restrição de que a divergência de *Kullback-Leibler* (KL) média entre a

nova política e a anterior seja limitada por um limiar δ:

$$\max_{\theta} \quad \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t \right] \tag{II.9}$$

$$\text{sujeito a} \quad \hat{\mathbb{E}}_t \left[D_{\text{KL}} \left(\pi_{\theta_{\text{old}}}(\cdot|s_t) \,\|\, \pi_\theta(\cdot|s_t) \right) \right] \leq \delta \tag{II.10}$$

onde:

- $\pi_\theta(a_t|s_t)$ é a nova política parametrizada;
- $\pi_{\theta_{\text{old}}}(a_t|s_t)$ é a política anterior;
- $\hat{A}_t$ representa a estimativa da função vantagem no tempo t;
- D_{KL} é a divergência de Kullback-Leibler entre as distribuições de probabilidade;
- δ é um hiperparâmetro que define o limite da região de confiança.

Essa abordagem pode ser interpretada como uma tentativa de limitar o quanto a nova política pode se afastar da antiga em uma única atualização, evitando assim degradações bruscas de desempenho. A restrição de KL é tratada por aproximações numéricas eficientes, como o método de gradiente conjugado.

Na prática, uma versão alternativa da restrição pode ser utilizando uma penalização com termo de KL, resultando na função objetivo:

$$L(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t - \beta D_{\text{KL}} \left(\pi_{\theta_{\text{old}}}(\cdot|s_t) \,\|\, \pi_\theta(\cdot|s_t) \right) \right] \tag{II.11}$$

onde β é um coeficiente que controla o peso da penalidade. No entanto, essa formulação requer ajuste cuidadoso de β, o que pode dificultar a aplicabilidade do método em diferentes domínios.

II.2.7 Função Objetivo Substituta

A função objetivo substituta é uma formulação amplamente utilizada em métodos de gradiente de política para otimizar os parâmetros da política sem depender diretamente da função de valor-ação $Q(s, a)$. Essa abordagem visa estimar a direção de melhoria da política com base em trajetórias amostradas, sem a necessidade de avaliar diretamente a função de desempenho esperada do agente.

Dada uma política estocástica $\pi_\theta(a_t|s_t)$ parametrizada por θ, a função objetivo substituta visa maximizar o retorno esperado ponderado pela vantagem estimada. Sua forma básica, derivada do teorema do gradiente de política, é expressa por:

$$L_{\mathrm{PG}}(\theta) = \hat{\mathbb{E}}_t \left[\log \pi_\theta(a_t|s_t)\, \hat{A}_t \right] \tag{II.12}$$

onde:

- $\pi_\theta(a_t|s_t)$ representa a probabilidade de selecionar a ação a_t no estado s_t segundo a política atual;
- $\hat{A}_t$ é uma estimativa da função vantagem no tempo t;
- $\hat{\mathbb{E}}_t[\cdot]$ denota a média empírica sobre amostras obtidas da interação com o ambiente.

No entanto, esse objetivo puro tende a ser sensível a grandes atualizações dos parâmetros, o que pode degradar rapidamente a política. Para contornar esse problema, algoritmos como TRPO e PPO introduzem uma forma modificada da função objetivo, utilizando razões de probabilidade entre a nova e a antiga política.

A vantagem de utilizar a função objetivo substituta é permitir múltiplas atualizações sobre o mesmo conjunto de dados coletados, desde que se mantenha o controle sobre o desvio entre a política antiga e a nova. Essa estrutura é a base para o desenvolvimento de algoritmos mais robustos, como o PPO, que utiliza versões modificadas dessa função com técnicas de *clipping* ou penalidades de divergência KL, conforme será detalhado na Seção II.2.9.

II.2.8 Coeficiente de Penalidade KL Adaptável

Uma das estratégias consideradas para controlar a distância entre políticas sucessivas em algoritmos de otimização de política é a penalidade baseada na divergência de KL. Essa abordagem busca evitar atualizações muito agressivas da política, introduzindo um termo penalizador na função objetivo que cresce conforme a nova política π_θ se afasta da política anterior $\pi_{\theta_{\mathrm{old}}}$.

A função objetivo com penalidade de KL é expressa como:

$$L_{\mathrm{KLPEN}}(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)} \hat{A}_t - \beta\, D_{\mathrm{KL}} \left(\pi_{\theta_{\mathrm{old}}}(\cdot|s_t) \,\|\, \pi_\theta(\cdot|s_t) \right) \right] \tag{II.13}$$

onde:

- $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}$ é a razão de probabilidade;
- $\hat{A}_t$ é a vantagem estimada;

- D_{KL} é a divergência KL entre as distribuições de política antiga e nova;
- β é o coeficiente de penalidade que regula a influência da divergência na função objetivo.

Um dos principais desafios dessa abordagem é escolher um valor apropriado para β. Um valor pequeno pode permitir atualizações instáveis, enquanto um valor muito grande pode frear excessivamente o aprendizado. Para contornar esse problema, Schulman et al. [2017] propõem um esquema adaptativo para ajustar β dinamicamente a cada iteração, com base no valor atual da divergência KL.

Nesse esquema, quando a divergência média d se encontra substancialmente abaixo do valor-alvo d_{targ}, o coeficiente é reduzido, isto é, $\beta \leftarrow \beta/2$, favorecendo atualizações mais amplas. Se, ao contrário, d ultrapassa significativamente o valor esperado, o coeficiente é ampliado, de modo que $\beta \leftarrow \beta \times 2$, reforçando o controle sobre as mudanças na política. Nos casos em que a divergência permanece em um intervalo aceitável em torno do alvo, o valor de β é mantido, garantindo estabilidade no processo de otimização.

$$d = \hat{\mathbb{E}}_t \left[D_{\mathrm{KL}} \left(\pi_{\theta_{\mathrm{old}}}(\cdot|s_t) \,\|\, \pi_\theta(\cdot|s_t) \right) \right] \tag{II.14}$$

Essa técnica de ajuste adaptativo permite que o algoritmo mantenha a divergência KL em torno de um valor ideal, proporcionando uma atualização mais segura e eficiente da política sem depender de ajustes manuais intensivos dos hiperparâmetros. Apesar disso, os autores observam que, na prática, o uso do objetivo com *clipping* (ver Seção II.2.9) tende a produzir melhores resultados de maneira mais simples, tornando o KL adaptativo mais útil como alternativa ou mecanismo complementar.

II.2.9 Proximal Policy Optimization (PPO)

O PPO é uma evolução dos métodos baseados em políticas, oferecendo uma abordagem mais estável e eficiente para o treinamento de agentes em ambientes com alta dimensionalidade de estados e ações, como os mercados financeiros.

O PPO busca otimizar uma política estocástica por meio de atualizações iterativas, mas controladas, evitando que a nova política se afaste excessivamente da anterior — um problema comum em métodos como o *REINFORCE* ou o *A3C*. Para isso, ele utiliza uma função de objetivo com um limite (*clipping*) que restringe a magnitude da atualização da política, promovendo um aprendizado mais estável.

No contexto deste trabalho, o PPO permite que o agente aprenda, de forma eficiente, a política ótima de negociação ao interagir com um ambiente de mercado simulado. A cada episódio, o

agente observa o estado do ambiente, que inclui variáveis de mercado e dados internos da carteira, escolhe ações (como comprar, vender ou manter uma posição), recebe recompensas baseadas em seu desempenho financeiro e atualiza sua política conforme os retornos obtidos. Ao longo do tempo, o PPO ajusta seus parâmetros para maximizar a função de retorno esperado, buscando estratégias que apresentem maior consistência e lucratividade frente aos desafios de mercados voláteis e não estacionários.

Por sua capacidade de equilibrar exploração e estabilidade no aprendizado, oferece um caminho promissor para a construção de agentes de negociação autônomos e robustos.

Capítulo III Trabalhos Relacionados

Neste capítulo são discutidos os trabalhos relacionados com aplicação de aprendizado por reforço para previsão no mercado financeiro. Para a realização da revisão sistemática da literatura, adotou-se a metodologia PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) [Page et al., 2021], a fim de garantir rigor e transparência no processo de seleção e análise dos estudos.

A estrutura do capítulo está organizada em duas seções principais. Na Seção III.1, detalhamos as etapas do processo de busca e seleção das publicações e descrevemos os critérios aplicados para determinar a elegibilidade e priorização dos estudos analisados; e na Seção III.2, apresentamos os resultados da revisão, destacando os principais trabalhos selecionados.

III.1 Seleção dos Trabalhos Relacionados

Para identificar os trabalhos relacionados, foi criada uma chave de busca, em inglês, composta por palavras-chave relacionadas ao tema do trabalho e seus sinônimos, visando ampliar o alcance da pesquisa. Assim, os descritores principais utilizados foram “Reinforcement Learning” (referente ao algoritmo) e “Financial Market”/“day trading”/“Day Trading” (referente à área de atuação). A tabela III.1 traz as chaves de busca utilizadas.

Tabela III.1: Chaves de busca utilizadas para encontrar os trabalhos relacionados neste trabalho. Pesquisa feita em 03/08/2024.

Palavras-chave	Quantidade de artigos encontrados
“Reinforcement Learning” AND “Financial Market’ ’	813
“Reinforcement Learning” AND “Day Trading’ ’	8
“Reinforcement Learning” AND “Day Trade’ ’	3

No escopo, foram considerados artigos em inglês, sem restrição temporal.

Para a busca dos trabalhos, foram selecionadas duas bases de dados: Scopus Elsevier [Elsevier, 2020]. A base Scopus Elsevier foi escolhida pela abundância de material científico disponível em seu acervo, que também abrange a área de pesquisa deste trabalho, mercado financeiro.

III.1.1 Critérios de Inclusão

Os critérios de inclusão têm a função de orientar o tema escolhido e excluir pesquisas que não atendem aos objetivos do estudo.

São três critérios, eles são:

- Critério 1: Trabalhos que abordem o uso de AR no mercado financeiro especificamente para estratégias de negociação intradiária;
- Critério 2: Trabalhos que abordem o uso de AR no mercado financeiro em outros contextos que não sejam negociação intradiária;
- Critério 3: Trabalhos que abordem métodos de AM e/ou Aprendizado Profundo (AP) no mercado financeiro, excetuando-se técnicas de AR, já contempladas nos critérios anteriores.

III.1.2 Critérios de Exclusão

Os critérios de exclusão têm a função de excluir trabalhos que não satisfazem este estudo.

São quatro critérios, eles são:

- Critério 1: Trabalhos que não tratem do mercado financeiro;
- Critério 2: Trabalhos que tratem do mercado financeiro, mas sem relação com AR, AM ou AP;
- Critério 3: Trabalhos que não oferecem informações relevantes utilizadas na pesquisa realizada;
- Critério 4: Trabalhos mais antigos (anteriores a 2020) e/ou menos citados em outros trabalhos.

III.1.3 Critérios de Priorização

Após a inclusão dos artigos na busca, são adotados critérios de prioridade que determinam a relevância do material obtido. Os artigos foram divididos em três níveis de prioridade:

- Prioridade Alta: Trabalhos relacionados ao uso de AR ou AM no mercado financeiro, focado em predições em negociação intradiária;
- Prioridade Média: Trabalhos relacionados ao uso de AR no mercado financeiro, focado em outros tipos de transações;
- Prioridade Baixa: Trabalhos relacionados ao uso de AM no mercado financeiro, excetuando-se técnicas de AR, focado em outros tipos de transações;

III.1.4 Resultados

O resultado da seleção dos trabalhos, obtidos utilizando a metodologia PRISMA, pode ser analisado conforme mostrado na Figura III.1. A quantidade de artigos identificados por meio da chave de busca ajustada com as palavras-chave selecionadas foi determinada na execução da busca em 03/08/2024. Foi feita uma atualização no dia 20/11/2025.

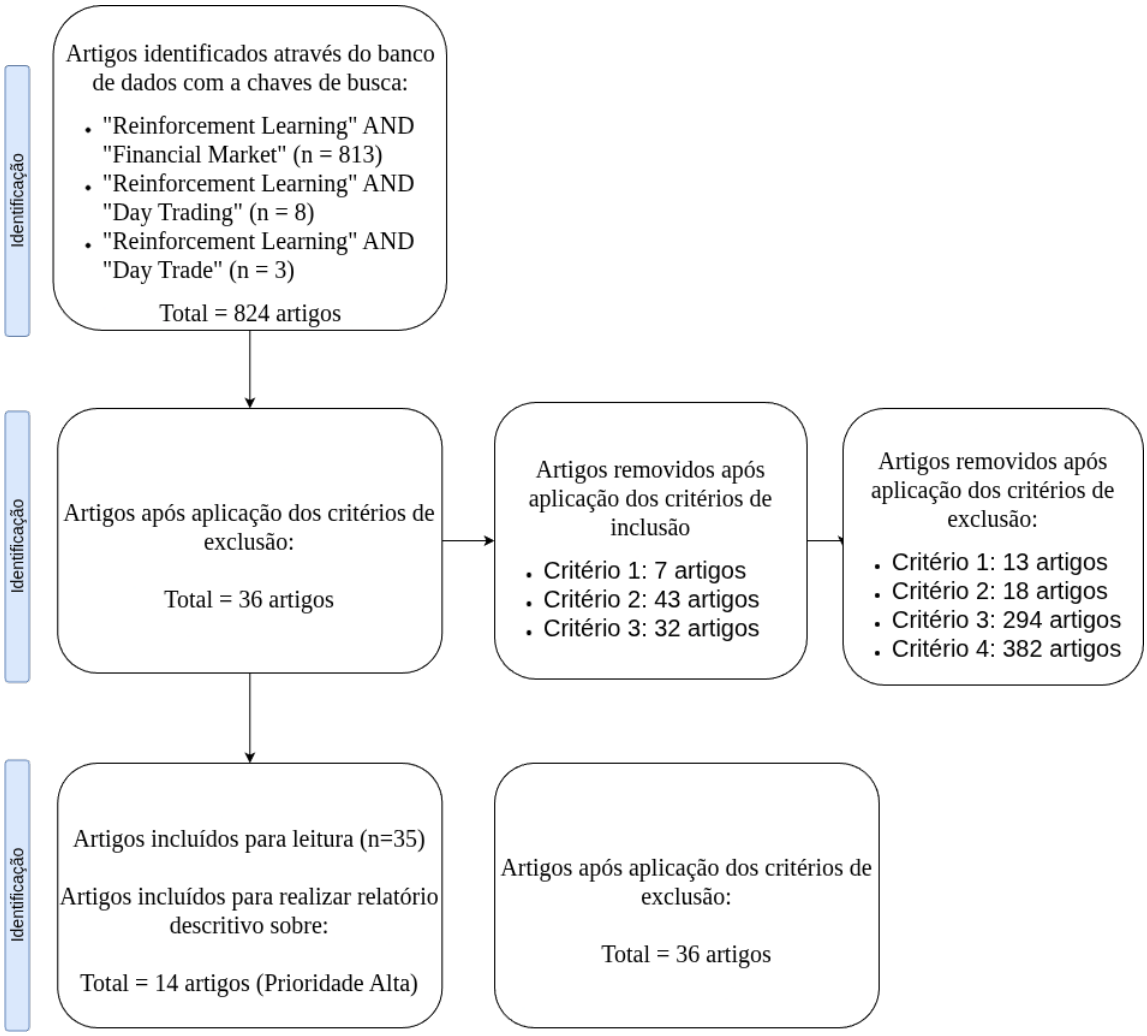


Figura III.1: Fluxograma feito com a metodologia PRISMA para seleção dos trabalhos relacionados

III.2 Síntese Comparativa

Esta seção apresenta uma síntese comparativa detalhada dos trabalhos analisados na revisão sistemática, com foco em compreender como os trabalhos das Tabelas III.2 e III.3 tem utilizado o ARP aplicados ao mercado financeiro, quais algoritmos são mais adotados, em quais contextos temporais e estruturais atuam melhor, e de que forma lidam com aspectos essenciais da microestrutura de mercado, como granularidade, ativos, ações contínuas e custos reais de execução.

O objetivo é estabelecer um posicionamento científico claro da proposta desta dissertação frente ao estado da arte, destacando explicitamente quais limitações recorrentes na literatura são mitigadas ou superadas pelo arcabouço desenvolvido neste trabalho.

Tabela III.2: Resumo dos principais trabalhos relacionados aplicando aprendizado por reforço em mercados financeiros, com foco no problema abordado, solução proposta e limitações observadas.

Autor	**Algoritmo**	**Problema abordado**	**Solução adotada**	**Limitação**
Conegundes and Pereira [2020]	DDPG	Desenvolver estratégia para negociação de ações brasileiras em negociação intradiária	Aplicação de DDPG com diferentes combinações de janelas deslizantes de dias anteriores.	Não valida a hipótese de janelas temporais mais curtas e utiliza apenas dados de abertura e fechamento dos leilões.
Qiu et al. [2021]	QF-TraderNet	Operar dados financeiros em categorias diversas com comportamentos distintos e robustos.	Modelo híbrido unindo finanças quânticas com Q-learning.	Alta complexidade computacional e utilização de dados apenas de abertura e fechamento dos leilões
Sattarov et al. [2020]	Deep NN	Aprender estratégia de negociação em criptomoedas altamente voláteis.	Rede neural profunda para aprender política a partir de dados de BTC, LTC e ETH.	Risco elevado de overfitting; volatilidade extrema pode reduzir estabilidade.
Altuner and Kilimci [2022]	DQN	Aprender a tendência das ações.	Análise de sentimentos de comentários financeiros do X, com inclusão de informação em um algoritmo de DQN.	Utiliza apenas dados de abertura e fechamento dos leilões e previsão apenas de tendência
Carta et al. [2021]	DQN	Capturar interação competitiva entre agentes no mercado.	Ambiente de múltiplos agentes DQNs simulando dinâmicas de negociação.	Cenário simplificado. Nem todas as dinâmicas reais são representadas.
Wu et al. [2023]	Triple Q-learning	Reduzir viés de superestimação em Q-learning.	Utilização de três estimadores para maior estabilidade nas decisões.	Custo computacional maior e desafios de escalabilidade.
Ghotbi and Zahedi [2024]	TDQNK	Negociar ativos diversos com foco em risco e retorno.	Integra funções de kernel ao Q-learning para maior generalização.	Alta exigência computacional e tuning complexo do kernel.
Wang et al. [2023]	DQN + LightGBM	Combinar AR com modelo tradicional para melhorar precisão.	Integração de DQN com LightGBM para geração de sinais híbridos.	Não demonstram como acharam algumas informações e grade dependência do modelo auxiliar.

Tabela III.3: Resumo dos principais trabalhos relacionados aplicando aprendizado por reforço em mercados financeiros, com foco no problema abordado, solução proposta e limitações observadas.

Autor	**Algoritmo**	**Problema abordado**	**Solução adotada**	**Limitação**
Jang and Seong [2023]	DDPG + Teoria do portfólio	Gerenciar portfólio maximizando risco-retorno.	Combina DDPG com princípios da teoria moderna de portfólio.	Sensível à composição do portfólio e dados históricos.
Cui et al. [2023]	PPO + CVaR	Incorporar controle de risco extremo em cripto.	Modelo PPO com otimização CVaR para priorizar risco extremo.	Modelo mais sensível e difícil de calibrar com alta variância.
Khan et al. [2023]	DQN + CNN	Aprender padrões temporais diretamente dos preços.	CNN para extrair features automaticamente e DQN para decisão.	Dependência de janelas temporais fixas e custo computacional elevado.
Zhang et al. [2023]	Adaptive RL (HFT)	Adaptar política rapidamente em ambiente HFT.	Modelo com adaptação contínua em alta frequência.	Naõ específica qual são os modelos utilizados como benchmarks.
Théate and Ernst [2021]	TDQN	Melhorar estabilidade e robustez do DQN.	Double DQN com priorização temporal de experiências.	Maior complexidade e dependência de hiperparâmetros.
Madhulatha and Ghori [2025]	DQN	Aumentar a adaptabilidade e precisão na previsão de taxas de câmbio.	Modelo híbrido combinando LSTM para dependências temporais e DQN para ajuste adaptativo baseado em recompensas, incorporando ainda indicadores macroeconômicos.	Alta complexidade computacional, baixa generalização (dados apenas USD/INR) e desafios de implementação em tempo real.
Este trabalho	**PPO, A2C, DDPG, SAC**	**Criar um algoritmo capaz de transacionar em negociação intradiária para um portfólio de ações.**	**Aplicação do PPO em dados intradiários de um portfólio de ações, em que supere benchmarks como S&P500 e Nasdaq**	

A análise da literatura evidencia inicialmente que há uma predominância marcante de métodos baseados em valor, especialmente variantes do algoritmo Deep Q-Network (DQN), em aplicações de negociação algorítmica. Trabalhos como os de Altuner and Kilimci [2022], Khan et al. [2023], Wang et al. [2023],Théate and Ernst [2021] e Madhulatha and Ghori [2025] utilizam DQN, Double DQN ou adaptações orientadas à mitigação de sobrestimação e instabilidade, tais como Triple Q-Learning ou o algoritmo especializado TDQN. A popularidade desses métodos decorre, em grande medida, da simplicidade operacional e da adequação ao espaço de ações discretas, que caracteriza a maioria dos ambientes artificiais empregados nesses estudos. No entanto, a dependência de ações discretas limita substancialmente sua aplicabilidade em cenários reais, onde a tomada de decisão tende a ser contínua e integrada à dinâmica complexa da microestrutura do mercado. Em ambientes com alta estocasticidade, como mercados intradiários ou de alta frequência, o DQN também apresenta instabilidades significativas e depende fortemente de técnicas de estabilização adicionais, que nem sempre são suficientes.

Por outro lado, métodos baseados em política determinística, como o Deep Deterministic Policy Gradient (DDPG), aparecem com menor frequência, mas desempenham papel relevante em trabalhos que demandam ações contínuas. Exemplos incluem Conegundes and Pereira [2020] e Jang and Seong [2023], que exploram a capacidade do DDPG de lidar com alocações contínuas ou ponderação de portfólios. Contudo, tais estudos permanecem restritos a granularidades diárias ou a ambientes com baixa complexidade estrutural, e apresentam frequentemente elevada sensibilidade a hiperparâmetros e dificuldade de generalização. Problemas de instabilidade, tendência ao overfitting e limitada adaptabilidade a mudanças abruptas de regime figuram como limitações recorrentes.

No que diz respeito a métodos mais robustos baseados em política proximal, observa-se uma presença ainda inferior na literatura. O algoritmo Proximal Policy Optimization (PPO), amplamente reconhecido pela estabilidade em espaços contínuos e por sua eficácia em cenários de forte variabilidade estocástica, aparece de maneira destacada apenas no trabalho de Cui et al. [2023]. Entretanto, mesmo nesse caso, o ambiente de aplicação é simplificado, de granularidade diária, restrito a poucos ativos e sem qualquer integração com dados de microestrutura ou fluxo de ordens. Ademais, trata-se de um contexto voltado a criptomoedas, cujo comportamento e estrutura operacional diferem substancialmente do mercado acionário intradiário. Assim, embora o PPO seja o algoritmo mais promissor em termos de estabilidade, sua adoção ainda é incipiente nas pesquisas analisadas.

Ao mapear os tipos de mercado e granularidades temporais utilizadas, nota-se um padrão: a grande maioria dos trabalhos opera em frequência diária, como demonstram Conegundes and

Pereira [2020], Jang and Seong [2023], Cui et al. [2023], Khan et al. [2023] e Théate and Ernst [2021]. Alguns estudos utilizam granularidade horária, como Sattarov et al. [2020], enquanto poucos avançam para níveis inferiores. Carta et al. [2021] empregam múltiplas resoluções, mas ainda assim sem utilizar microestrutura real. O único estudo que se aproxima de granularidades finas é o de Zhang et al. [2023], que opera com intervalos de segundos. Contudo, mesmo esse trabalho apresenta limitações relevantes, como ausência de transparência metodológica, falta de integração com dados reais do livro de ofertas, ausência de políticas contínuas. Assim, apesar de indicar um movimento em direção a ambientes mais realistas, a literatura continua distante de empregá-los de forma consistente.

Outro aspecto refere-se ao uso de múltiplos ativos. Embora alguns estudos abordem conjuntos maiores, como Carta et al. [2021], Jang and Seong [2023] e Cui et al. [2023], a maioria trabalha com um único ativo ou com portfólios reduzidos, frequentemente construídos a partir de dados diários. Mesmo nos poucos trabalhos multi-ativos, há ausência de integração com dados intradiários de microestrutura, o que dificulta a transposição dos resultados para cenários reais. A capacidade de operar simultaneamente múltiplos ativos com granularidade fina representa, portanto, uma lacuna significativa na literatura.

No que tange ao uso de dados reais de microestrutura, a literatura mostra uma limitação ainda mais evidente. Com exceção parcial de Zhang et al. [2023], nenhum estudo utiliza o livro de ofertas com múltiplos níveis. A maioria baseia-se exclusivamente em dados agregados, incapazes de capturar o comportamento dinâmico das ordens, a profundidade do mercado, a liquidez instantânea e as interações entre posições relativas de compra e venda. Essa limitação compromete a validade de modelos que buscam se aproximar do comportamento intradiário real, uma vez que decisões eficazes nesse nível dependem diretamente da compreensão do fluxo de ordens, da volatilidade de curtíssimo prazo e da disponibilidade de liquidez em múltiplos níveis.

A análise das limitações metodológicas também revela que a maioria dos estudos não incorpora custos de transação, latência, impacto no preço ou *slippage*. Essa ausência, frequentemente justificada pela simplificação dos ambientes de simulação, compromete a aplicabilidade prática dos modelos propostos, especialmente em mercados intradiários altamente competitivos. A negligência desse conjunto de fatores pode inflar artificialmente os resultados de desempenho e mascarar fragilidades que só se revelariam em ambientes reais.

Tendo em vista esse panorama, torna-se evidente a contribuição desta dissertação. A proposta desenvolvida aqui avança significativamente o estado da arte ao empregar dados de granularidade minuto a minuto, extraídos diretamente da CBOE, com integração completa dos dez níveis do livro de ofertas, tanto do lado comprador quanto vendedor, para múltiplas ações transacionadas

simultaneamente. Nenhum dos trabalhos analisados reúne simultaneamente granularidade fina, dados reais de microestrutura, política contínua e ambiente multi-ativo. Além disso, o uso de ações contínuas no intervalo real $[-1, 1]$ aproxima substancialmente o agente do comportamento real de negociação, permitindo variação proporcional de posição em vez de decisões discretas simplificadas. A adoção do algoritmo PPO, reconhecidamente mais estável e adequado para cenários contínuos, reforça a robustez do modelo, sobretudo em um ambiente intradiário complexo e altamente estocástico.

A presente dissertação também se diferencia por modelar slippage e latência de forma explícita, simulando um ambiente coerente com a lógica operacional empregada por plataformas reais de negociação, aproximando-se de forma inédita da estrutura funcional da própria CBOE. Ao incorporar esses elementos, o modelo desenvolvido não apenas mitiga limitações centrais observadas na literatura, como também aumenta substancialmente a utilidade prática dos resultados e sua capacidade de generalização para cenários de produção.

Capítulo IV Conjuntos de Dados

Neste capítulo, apresentamos o processo de construção da base de dados e da representação do ambiente utilizadas neste trabalho, abrangendo desde a obtenção das informações intradiárias até a análise exploratória dos dados. A Seção IV.1 apresenta a arquitetura e o funcionamento do sistema de *web scraping* desenvolvido para coletar dados intradiários do CBOE. Na Seção IV.2, discutimos as etapas de tratamento e limpeza dos dados, com foco em garantir consistência e qualidade das informações. A Seção IV.3 trata da construção das entradas do agente, descrevendo a seleção das variáveis, os procedimentos de normalização, a agregação temporal e a representação final do estado utilizado pelo modelo de Aprendizado por Reforço Profundo (ARP). Por fim, a Seção IV.4 apresenta uma análise exploratória dos dados coletados, incluindo estudos de retorno percentual acumulado, valorização de portfólios e análise de tendências, fornecendo uma visão geral das características estatísticas e dinâmicas do conjunto de dados empregado nos experimentos subsequentes.

IV.1 Obtenção dos dados

A etapa de obtenção dos dados consistiu na implementação de um sistema de *web scraping* responsável por extrair, em tempo real, informações intradiárias disponibilizadas pelo CBOE. A arquitetura desse sistema, ilustrada na Figura IV.1, foi projetada para realizar coletas contínuas durante o horário das 10h às 18h, em dias úteis de operação da bolsa de valores norte-americana.

O processo de coleta foi implementado em Python e opera da seguinte forma: o script realiza requisições sucessivas às páginas individuais de cada companhia monitorada, aguarda-se um intervalo de dois segundos, a fim de assegurar carregamento completo do conteúdo e, em seguida, extrai os elementos relevantes do arquivo *.html*. Esse procedimento é executado para cada uma das 19 empresas selecionadas para o estudo: Apple, Tesla, Cisco, Microsoft, General Electric, Ford, Citigroup, Freeport-McMoRan, Bank of America, Coca-Cola, Intel, General Motors, American Airlines, Norwegian Cruise Line, J.P. Morgan, Pfizer, Morgan Stanley, Delta Air Lines e Newmont Corp.

Cada registro coletado recebe adicionalmente um carimbo temporal obtido diretamente da máquina responsável pela coleta, garantindo o alinhamento cronológico das observações. Após

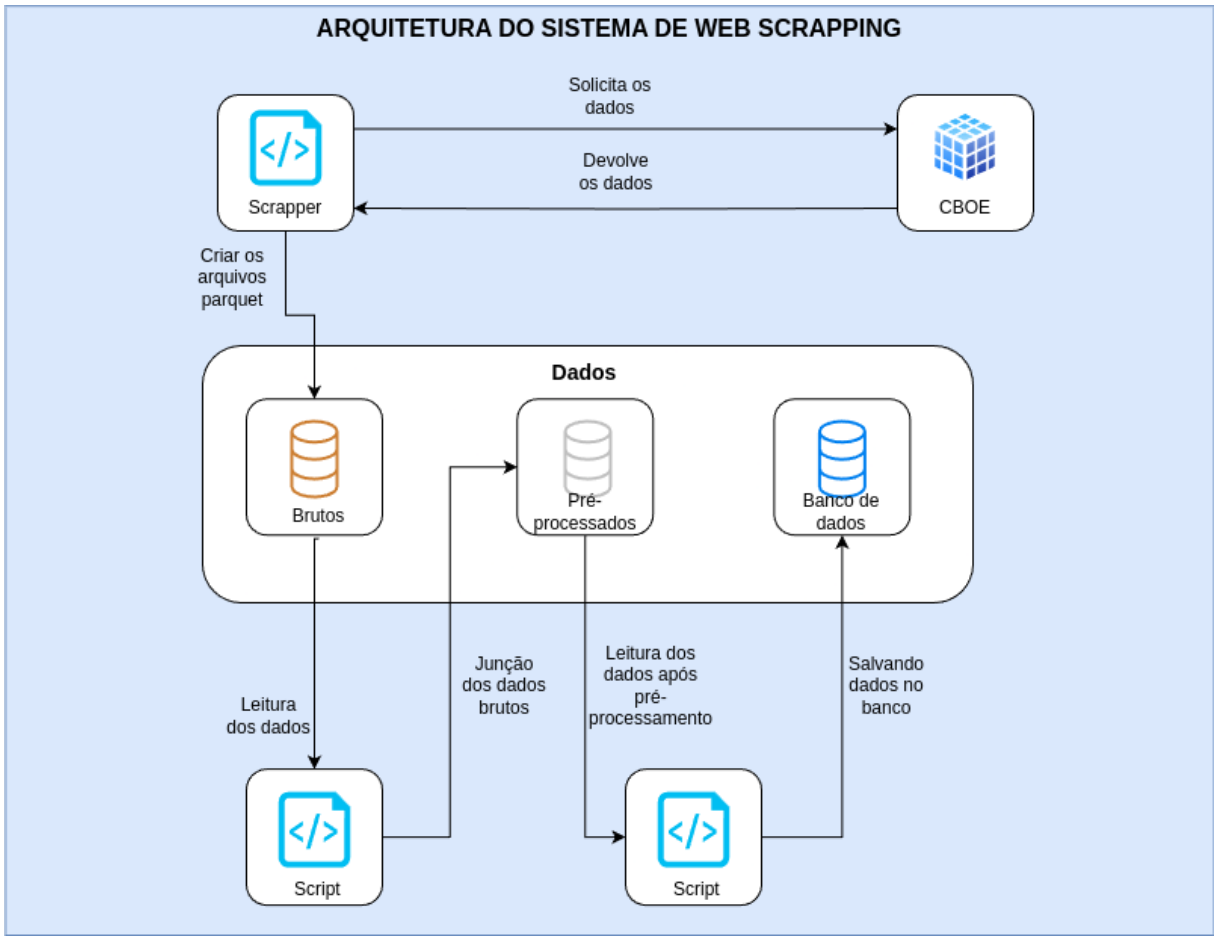


Figura IV.1: Arquitetura do sistema de coleta de dados implementado para extrair informações intradiárias do CBOE. A figura ilustra o fluxo completo do processo: envio das requisições às páginas individuais de cada companhia, aguardo para carregamento do conteúdo, extração estruturada dos elementos do arquivo `.html`, registro do carimbo temporal local e armazenamento dos dados tanto no SGBD MySQL quanto em arquivos no formato *parquet* (camadas *raw* e *preprocessed*).

concluída a extração diária, os dados são armazenados em duas estruturas: (i) no SGBD MySQL, para facilitar consultas estruturadas, e (ii) em arquivos *parquet*, salvos tanto em formato bruto (camada *raw*) quanto em forma consolidada (camada *preprocessed*).

IV.2 Tratamento de dados

Após a coleta, foi realizada a etapa de tratamento dos dados, com foco em garantir consistência, padronização e qualidade dos dados que serão utilizadas no treinamento e na avaliação do agente.

Inicialmente, todos os campos numéricos foram convertidos para formatos padronizados, incluindo a normalização de separadores decimais e a harmonização de tipos numéricos. Em seguida, conduziu-se uma análise exploratória da integridade dos dados. Verificou-se que aproxi-



AAPL | Orders Accepted | Total Volume

APPLE INC COM | 374,134 | 319,412

	TOP OF BOOK			LAST 10 TRADES	
	Shares	Price	Time	Price	Shares
ASKS	37	268.87	10:17:09	268.77	88
	37	268.86	10:17:06	268.76	10
	39	268.85	10:17:05	268.76	2
	14	268.84	10:17:05	268.76	6
	24	268.80	10:17:04	268.74	6
BIDS	10	268.76	10:17:04	268.75	30
	10	268.75	10:17:04	268.75	100
	35	268.73	10:17:04	268.75	100
	10	268.71	10:17:03	268.72	20
	39	268.70	10:17:03	268.72	15

Figura IV.2: Dados intradiários da Apple obtidos no site da CBOE. Esses dados incluem preços, quantidades negociadas e demais informações de mercado que serviram de insumo para a modelagem do ambiente de negociação e para os experimentos apresentados neste trabalho.

madamente 4% das observações apresentavam valores faltantes, ocasionados por falta de informações no início e fim do dia para as ações, lentidão no carregamento das páginas ou interrupções momentâneas no serviço do CBOE. Esses registros foram removidos para evitar ruído nas etapas posteriores. Adicionalmente, identificou-se que cerca de 7% das linhas eram duplicadas, resultado de atualizações mínimas no intervalo de tempo entre coletas, e foram eliminadas.

As Tabelas IV.1 e IV.2 demonstram uma amostragem dos dados após o tratamento.

Data	**Hora**	**Ações - Livro de ofertas**	**Preços - Livro de ofertas**	**Preço por Ação - Últimas 10 transações**	**Ações vendidas - Últimas 10 transações**	**Tipo de transação - Livro de ofertas**
2024-07-10	11:21:49	431,0	230,99	230,940	1,0	asks
2024-07-10	11:21:47	130,0	230,98	230,900	1,0	asks
2024-07-10	11:21:47	100,0	230,97	230,900	400,0	asks
2024-07-10	11:21:46	30,0	230,96	230,910	48,0	asks
2024-07-10	11:21:47	6,0	230,95	230,910	52,0	asks
2024-07-10	11:21:43	110,0	230,92	230,910	1,0	asks
2024-07-10	11:21:41	30,0	230,91	230,911	1,0	bids
2024-07-10	11:21:40	222,0	230,90	230,920	3,0	bids
2024-07-10	11:21:40	10,0	230,89	230,920	5,0	bids
2024-07-10	11:21:40	336,0	230,88	230,920	7,0	bids

Tabela IV.1: Amostra dos dados coletados no banco de dados da empresa Apple, contendo informações de preço, quantidade disponível no livro de ofertas, volume negociado e tipo de transação (*asks* ou *bids*). Esses dados representam a estrutura informacional disponibilizada pela CBOE, que reflete as condições reais observadas pelo agente durante o processo de aprendizado e tomada de decisão.

Data	Hora	Ações - Livro de ofertas	Preços - Livro de ofertas	Preço por Ação - Últimas 10 transações	Ações vendidas - Últimas 10 transações	Tipo de transação - Livro de ofertas
2024-07-10	11:21:50	205,0	261,98	261,870	4,0	asks
2024-07-10	11:21:50	355,0	261,97	261,890	0,0	asks
2024-07-10	11:21:50	5,0	261,96	261,890	100,0	asks
2024-07-10	11:21:50	170,0	261,95	261,890	5,0	asks
2024-07-10	11:21:50	87,0	261,93	261,900	4,0	asks
2024-07-10	11:21:50	255,0	261,79	261,900	78,0	bids
2024-07-10	11:21:49	50,0	261,77	261,800	22,0	bids
2024-07-10	11:21:49	50,0	261,75	261,800	9,0	bids
2024-07-10	11:21:48	54,0	261,72	261.680	18,0	bids
2024-07-10	11:21:48	55,0	261,70	261.680	19,0	bids

Tabela IV.2: Amostra dos dados coletados no banco de dados da empresa Tesla, contendo informações de preço, quantidade disponível no livro de ofertas, volume negociado e tipo de transação (*asks* ou *bids*). Esses dados representam a estrutura informacional disponibilizada pela CBOE, que reflete as condições reais observadas pelo agente durante o processo de aprendizado e tomada de decisão.

IV.3 Construção das Entradas do Agente

A etapa de construção das entradas do agente estabelece a ponte entre os dados intradiários obtidos do CBOE e a representação vetorial consumida pelo modelo de Aprendizado por Reforço. Enquanto as seções anteriores trataram da coleta e do tratamento dos dados brutos, nesta seção descrevemos como essas informações são transformadas em observações estruturadas, formando o estado s_t utilizado pelo agente durante o treinamento e a avaliação.

Essa etapa é fundamental, pois determina como o ambiente é percebido pelo agente e influência diretamente sua capacidade de identificar padrões relevantes, avaliar riscos e tomar decisões lucrativas.

IV.3.1 Seleção das variáveis observadas

Cada observação temporal contém informações provenientes do livro de ofertas e das últimas negociações efetuadas na bolsa. Após o tratamento dos dados brutos, as seguintes variáveis foram selecionadas para compor a representação informacional do ambiente:

- **Cinco melhores preços de compra (*bid*) e respectivas quantidades**;
- **Cinco melhores preços de venda (*ask*) e respectivas quantidades**;
- **Preços das últimas dez negociações efetivamente realizadas**;
- **Quantidades negociadas nas últimas dez transações**;

- **Dados internos do agente**: saldo disponível, capital investido, quantidade de cotas em carteira.

Essas variáveis foram selecionadas por refletirem, simultaneamente, a microestrutura do mercado (por meio do livro de ofertas) e o comportamento recente dos agentes econômicos (por meio da série de transações).

IV.3.2 Normalização e padronização

Para assegurar que todas as variáveis apresentassem escalas comparáveis, foi aplicada uma padronização baseada em variações relativas. Esse procedimento reduz o impacto de diferenças de magnitude entre ativos e evita que o modelo atribua relevância indevida a preços absolutos mais elevados.

As seguintes transformações foram utilizadas:

- Preços e quantidades do livro de ofertas foram convertidos para *variações percentuais* em relação ao preço médio do minuto anterior;
- Preços das últimas negociações foram convertidos para *retornos logarítmicos*;
- Quantidades negociadas foram escaladas via normalização min-max;
- As variáveis internas do agente foram normalizadas por seus valores máximos possíveis ao longo do episódio.

Essa estratégia facilita o aprendizado e estabiliza o comportamento numérico da rede neural que parametriza a política e a função de valor.

IV.3.3 Agregação temporal

Para que o agente capturasse padrões de curto prazo, foi incorporada uma janela deslizante contendo as últimas μ observações. Assim, cada estado inclui não apenas a observação corrente, mas também um histórico recente:

$$s_t = \big[x_{t-\mu+1},\, x_{t-\mu+2}, \ldots, x_t\big],$$

em que cada vetor x_t contém todas as variáveis descritas anteriormente.

A profundidade temporal μ foi definida experimentalmente, conforme discutido no Capítulo VI, e buscou representar o equilíbrio entre capturar dinâmica temporal suficiente e evitar redundância excessiva.

IV.3.4 Representação final do estado

Após a seleção, normalização e agregação temporal, as variáveis são concatenadas em um único vetor:

$$x_t \in \mathbb{R}^d,$$

onde d depende do número de ativos considerados, das variáveis extraídas do livro de ofertas e da profundidade μ da janela temporal.

Assim, o estado final fornecido ao agente é um tensor:

$$s_t \in \mathbb{R}^{\mu \times d},$$

representando o conjunto completo de informações disponíveis no instante t. Esse tensor é utilizado como entrada da política $\pi(a \mid s)$ durante o processo de tomada de decisão descrito na Seção V.2.

A construção rigorosa desse vetor de entrada assegura que o agente disponha de uma representação rica e coerente da dinâmica intradiária, permitindo que o modelo aprenda estratégias sensíveis à microestrutura do mercado e às condições recentes de negociação.

A Tabela IV.3 sintetiza todas as variáveis que compõem as observações x_t utilizadas na formação do estado s_t do agente. Ela contempla informações provenientes do livro de ofertas, incluindo os cinco melhores preços e quantidades de compra e venda, bem como os preços e volumes das últimas dez negociações registradas, oferecendo ao agente uma visão detalhada da microestrutura do mercado. Além dos dados brutos, a tabela inclui variáveis derivadas, como o retorno logarítmico e a variação percentual do preço médio, que ajudam a capturar dinâmicas de curto prazo. Também são incorporadas variáveis internas relacionadas à posição do agente, como saldo disponível, valor investido, número de cotas em carteira, valor total do portfólio e lucro ou prejuízo não realizado, essenciais para a tomada de decisão financeira. Por fim, a tabela destaca que o estado inclui uma estrutura temporal baseada em uma janela de μ observações históricas, garantindo que o agente tenha acesso a informações sequenciais necessárias para identificar padrões temporais relevantes.

IV.4 Análise exploratória

Nesta seção, são apresentadas análises referentes aos dados utilizados como base para o treinamento e teste do agente. Todos os dados empregados nos processos de treino e teste dizem respeito as empresas listadas na Seção IV.1. Na Tabela IV.4 reúne as empresas analisadas, o

Tabela IV.3: Conjunto de variáveis utilizadas na construção das observações x_t e, consequentemente, do estado s_t. As variáveis provêm do livro de ofertas, das últimas negociações efetivadas e de informações internas ao agente.

Nome da Variável	Descrição
Livro de ofertas (Top-of-Book)	
Bid_Price_[1..5]	Cinco melhores preços de compra disponíveis no livro de ofertas no instante t.
Bid_Volume_[1..5]	Quantidade disponível para compra em cada um dos cinco níveis de melhor oferta.
Ask_Price_[1..5]	Cinco melhores preços de venda ofertados no instante t.
Ask_Volume_[1..5]	Quantidade disponível para venda em cada um dos cinco níveis de melhor oferta.
Últimas transações	
LastTrade_Price_[1..10]	Preços das últimas dez negociações efetivamente realizadas.
LastTrade_Volume_[1..10]	Quantidade negociada em cada uma das últimas dez transações.
Variáveis derivadas	
Return_Log	Retorno logarítmico do preço médio entre $t-1$ e t.
Price_Change	Variação percentual do preço médio: $\frac{\text{Mid}_t - \text{Mid}_{t-1}}{\text{Mid}_{t-1}}$.
Variáveis internas do agente	
Cash_Balance	Saldo disponível para novas operações no instante t.
Invested_Value	Valor investido na posição corrente.
Shares_Held	Quantidade de cotas em carteira.
Portfolio_Value	Valor total do portfólio: posição + saldo.
Unrealized_PnL	Lucro ou prejuízo não realizado com base no preço atual.
Estrutura temporal	
j-ésima observação histórica $s_t = [x_{t-\mu+1}, \ldots, x_t]$.	Cada uma das μ observações anteriores incluídas na janela temporal:

período analisado, a quantidade de observações registradas em cada uma delas, além da referência às respectivas tabelas de estatísticas descritivas, que detalham as principais características dos dados coletados em cada período.

IV.4.1 Análise de Retorno Percentual Acumulado

O retorno percentual acumulado é uma métrica utilizada para avaliar a valorização de um ativo financeiro ao longo do tempo. Ele indica a variação percentual do preço de uma ação em relação ao seu valor inicial em um determinado período de observação, permitindo uma comparação direta do desempenho entre diferentes ativos, independentemente dos seus preços absolutos iniciais.

Essa abordagem é particularmente útil para destacar tendências de crescimento ou queda, bem como para facilitar análises comparativas em estudos de portfólios, estratégias de investimento ou avaliação de risco entre ativos com escalas diferentes de preços.

Tabela IV.4: Empresas analisadas no período de 03/03/2025 a 31/07/2025. As empresas possuem, cada uma, 432.160 observações.

Empresa	**Ramo**	**Sumário Estatístico**
American Airlines	Aéreas	Tabela A.3
Apple	Tecnologia	Tabela A.1
Bank of America	Financeiras	Tabela A.2
Cisco	Tecnologia	Tabela A.1
Citigroup	Financeiras	Tabela A.2
Coca-Cola	Alimentação	Tabela A.6
Delta Air Lines	Aéreas	Tabela A.3
Ford	Automóveis	Tabela A.4
Freeport-McMoRan	Mineradoras	Tabela A.5
General Electric	Diversos	Tabela A.6
General Motors	Automóveis	Tabela A.4
Intel	Tecnologia	Tabela A.1
JPMorgan Chase	Financeiras	Tabela A.2
Microsoft	Tecnologia	Tabela A.1
Morgan Stanley	Financeiras	Tabela A.2
Newmont Corp	Mineradoras	Tabela A.5
Norwegian Cruise	Setores Diversos	Tabela A.6
Pfizer	Farmacêutica	Tabela A.6
Tesla	Automóveis	Tabela A.4

Neste trabalho, adotamos a normalização do preço inicial das ações. A partir desse ponto de referência, as variações nos preços subsequentes são expressas como múltiplos ou frações desse valor inicial. Com isso, é possível analisar a trajetória de valorização relativa das empresas, de forma padronizada.

IV.4.1.1 Empresas de um mesmo setor

A Figura IV.3 apresenta os gráficos do retorno percentual acumulado para quatro empresas do setor de tecnologia (Microsoft, Apple, Cisco e Intel) no período analisado. Observa-se que, apesar de diferentes comportamentos individuais, é possível comparar diretamente o desempenho proporcional de cada empresa ao longo do tempo. Observa-se que a única empresa que teve valorização no período foi a Microsoft, sendo a Intel a com a maior desvalorização. Quando analisamos suas estatísticas na Tabela A.1, há uma grande diferença no preço dessas ações, tornando o retorno percentual uma métrica ideal para esse tipo de comparação entre valores tão distintos.

A Figura IV.4 apresenta os gráficos do retorno percentual acumulado para quatro empresas do setor de finanças (Bank of America, Citigroup, JP Morgan e Morgan Stanley) no período analisado. Observa-se que, exceto a Morgan Stanley, todas as empresas obtiveram um retorno positivo. O maior retorno do período foi da Citigroup, obtendo um retorno de quase 25%.

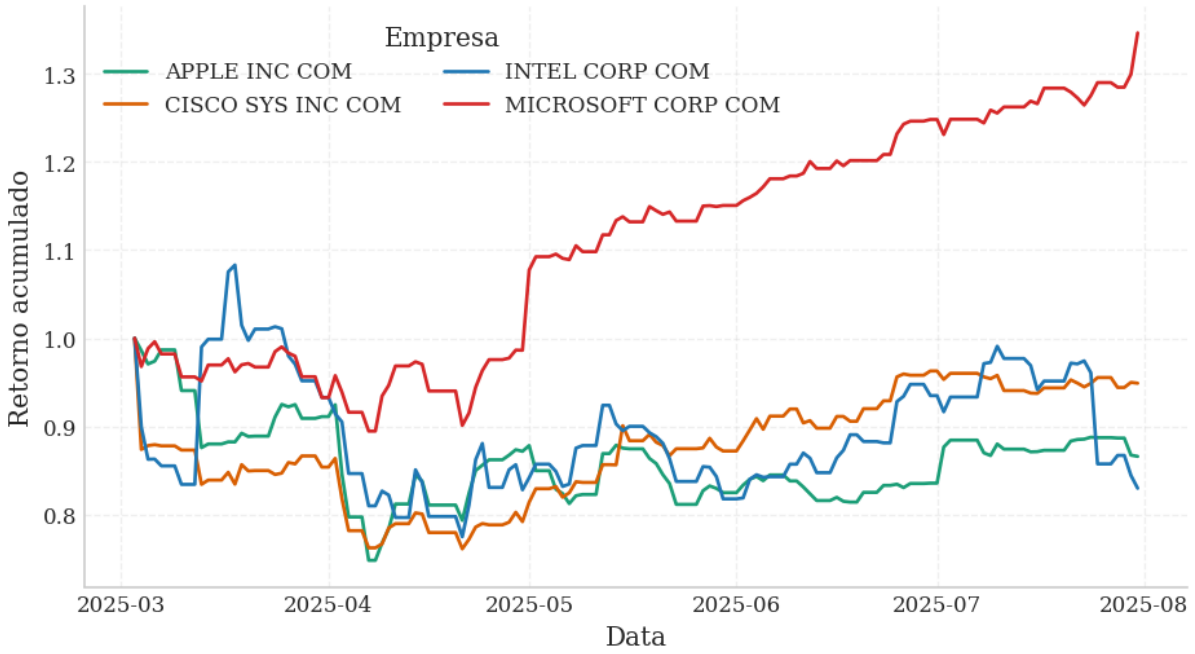


Figura IV.3: Gráfico comparando a evolução da valorização das empresas do setor de tecnologia ao longo do período de 03/03/2025 a 31/07/2025. A figura evidencia diferenças no comportamento de crescimento entre as companhias analisadas, permitindo observar tendências de desempenho relativo e períodos de maior volatilidade.

A Figura IV.5 apresenta os gráficos do retorno percentual acumulado para duas empresas do setor de linhas aéreas (American Airlines e Delta Air Lines) no período analisado. Observa-se que, o comportamento de retorno acumulado das empresas é bastante semelhante, demonstrando que as empresas desse setor valorizam e se desvalorizam com um padrão parecido. Ambas as empresas tiveram uma desvalorização no período analisado. Porém, em quase todos os dias, a American Airlines teve uma queda menor que a Delta Air Lines.

A Figura IV.6 apresenta os gráficos do retorno percentual acumulado para três empresas do setor automotivo (Tesla, Ford e GM) no período analisado. Observa-se que inicialmente há uma queda considerável da Tesla (carros elétricos) em relação as outras, que tem como maior foco carros a combustão. Entretanto, ao longo de maio há uma valorização da Tesla, terminando, no fim do período analisado, como a com a menor taxa de retorno. A Ford obtém um retorno de quase 15%.

A Figura IV.7 apresenta os gráficos do retorno percentual acumulado para duas empresas do setor de mineração (Freeport-Mcmoran, NewMont) no período analisado. Observa-se que apenas Newmont possui uma valorização, ultrapassando os 40%, sendo o maior retorno de todas as ações.

A Figura IV.8 apresenta os gráficos do retorno percentual acumulado para quatro empresas dos setores distintos (Coca-Cola, General Eletric, Pfizer e Norwegian Cruise) no período anali-

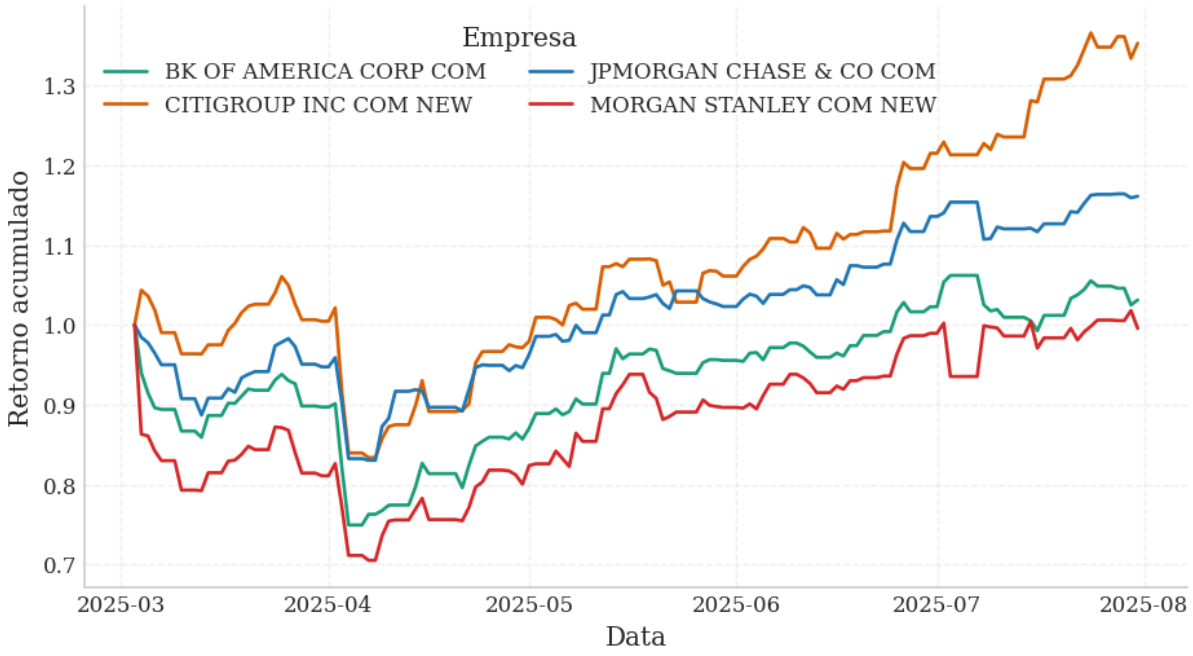


Figura IV.4: Gráfico apresentando a evolução da valorização das instituições financeiras ao longo do período de 03/03/2025 a 31/07/2025. A figura permite comparar o desempenho relativo entre as empresas do setor, destacando variações de crescimento, estabilidade e possíveis períodos de maior volatilidade.

sado. Observa-se que a General Eletric é a única empresa com um retorno positivo, sendo as outras três com retornos negativos. Norwegian Cruise tem a maior queda de todas as ações que analisamos.

IV.4.1.2 Valorização de portfólio

A valorização acumulada de um portfólio representa a evolução percentual do valor total de um conjunto de ativos ao longo do tempo, considerando uma base de comparação comum, normalmente o valor inicial do portfólio. Essa métrica permite avaliar o desempenho agregado de um grupo de ações, simulando o comportamento de um investidor que distribuiu igualmente seu capital entre os ativos do portfólio no início do período.

Na Figura IV.9, é possível observar a trajetória de valorização acumulada de um portfólio composto por empresas de setor distintos, que inclui as 19 companhias analisadas nesse trabalho. Para essa análise, os preços de cada ação foram normalizados em relação ao primeiro valor disponível no período. Assim, os valores subsequentes indicam a multiplicação do capital inicial.

Essa abordagem facilita a comparação entre ativos com diferentes faixas de preço e destaca a desempenho relativa de cada ação. Observa-se, na análise feita na Seção IV.4.1.1 ao longo do tempo, que algumas empresas apresentaram crescimento mais acentuado, enquanto outras mantiveram uma trajetória mais estável ou até declinante. A análise de retorno acumulado

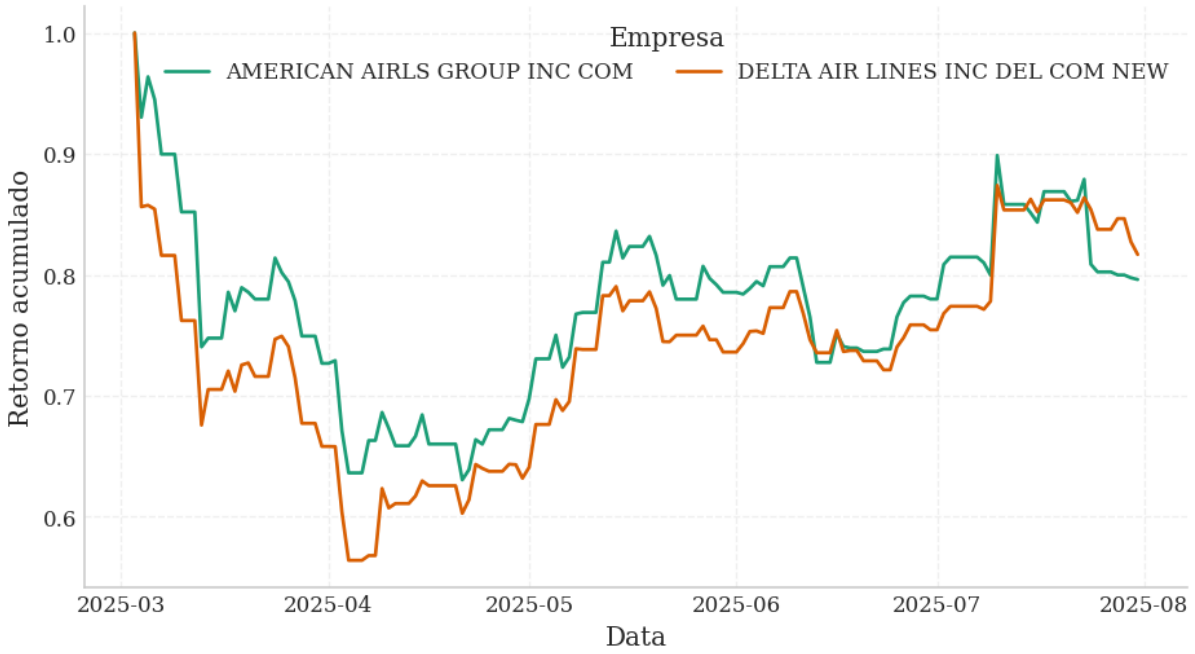


Figura IV.5: Gráfico mostrando a evolução da valorização das empresas do setor de linhas aéreas no período de 03/03/2025 a 31/07/2025. A figura permite comparar o desempenho relativo dessas companhias, evidenciando tendências de recuperação, variações sazonais e possíveis períodos de maior volatilidade característicos do setor.

dessa forma é útil para decisões estratégicas de alocação de recursos e avaliação de riscos em carteiras diversificadas.

IV.4.2 Análise de tendência

O teste de Mann-Kendall proposto por Mann [1945] e posteriormente aprimorado por Kendall [1975] e Gilbert [1987], é uma ferramenta estatística amplamente utilizada na análise de séries temporais. Seu principal objetivo é verificar se existe uma tendência monotônica significativa, ou seja, um comportamento consistentemente crescente ou decrescente, em uma determinada variável observada ao longo do tempo.

Uma tendência monotônica crescente indica que os valores da variável tendem a aumentar de forma sistemática, sem necessariamente obedecer a um padrão linear. Da mesma forma, uma tendência monotônica decrescente aponta para uma queda contínua nos valores ao longo do tempo. É importante destacar que o termo "monotônica" refere-se à consistência da direção da tendência, e não à sua intensidade ou forma exata.

O teste estatístico baseia-se na verificação da hipótese nula (H_0), que assume a inexistência de qualquer tendência temporal, ou seja, as observações são aleatórias e não seguem nenhum padrão direcional. A hipótese alternativa (H_1), por sua vez, postula a presença de uma tendência monotônica, seja ela ascendente ou descendente.

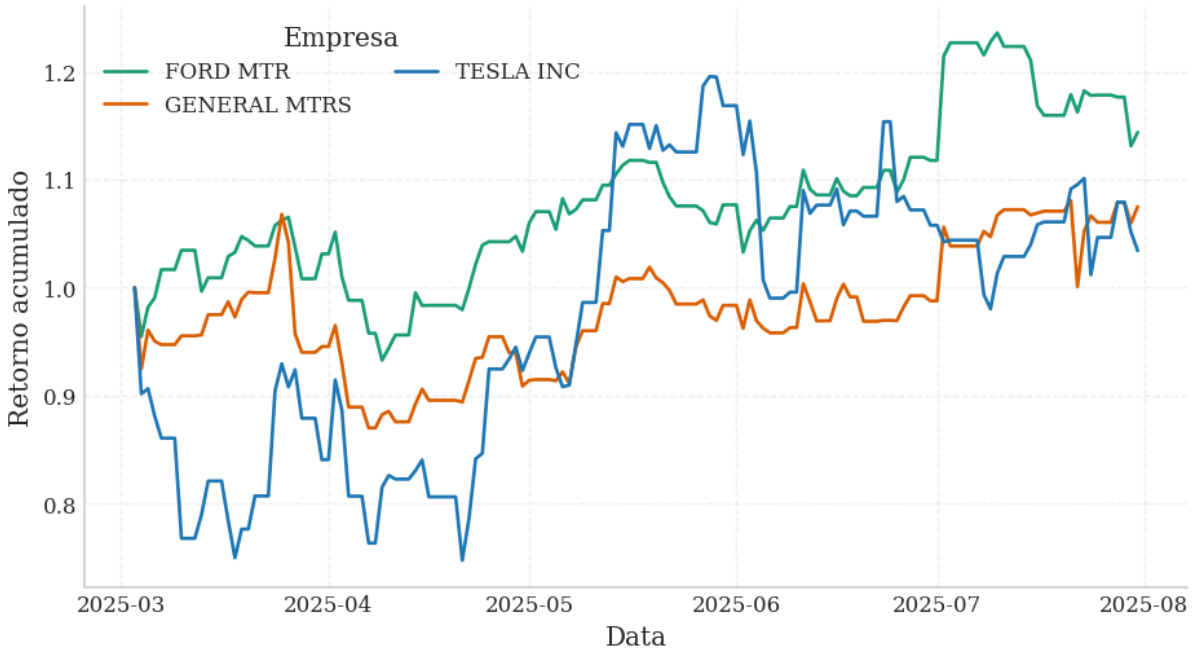


Figura IV.6: Gráfico apresentando a evolução da valorização das empresas do setor automotivo ao longo do período de 03/03/2025 a 31/07/2025. A figura possibilita comparar o desempenho relativo entre as companhias, destacando tendências de crescimento, oscilações de mercado e eventuais diferenças no comportamento financeiro dentro do setor.

Uma das principais vantagens do teste de Mann-Kendall em relação a outros métodos de análise de tendência é sua robustez estatística: ele não exige que os dados sigam uma distribuição normal (ou qualquer outra específica), tornando-o particularmente adequado para dados ambientais, econômicos ou financeiros que costumam apresentar assimetrias, valores anômalos ou valores ausentes. Aliás, o teste tolera valores faltantes, o que o torna aplicável mesmo em séries temporais incompletas Pandolfi et al. [2016].

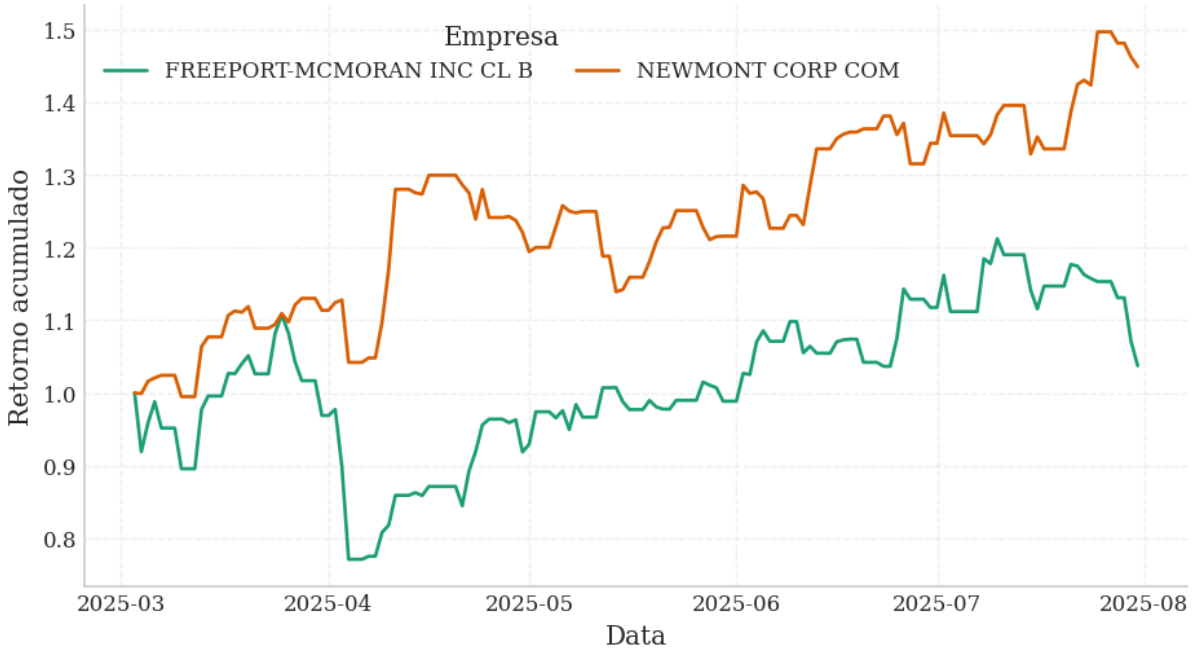


Figura IV.7: Gráfico apresentando a evolução da valorização das empresas do setor de mineração ao longo do período de 03/03/2025 a 31/07/2025. A figura permite comparar o desempenho relativo entre as companhias, destacando oscilações associadas às condições do mercado de commodities, variações de demanda e outros fatores que influenciam o setor.

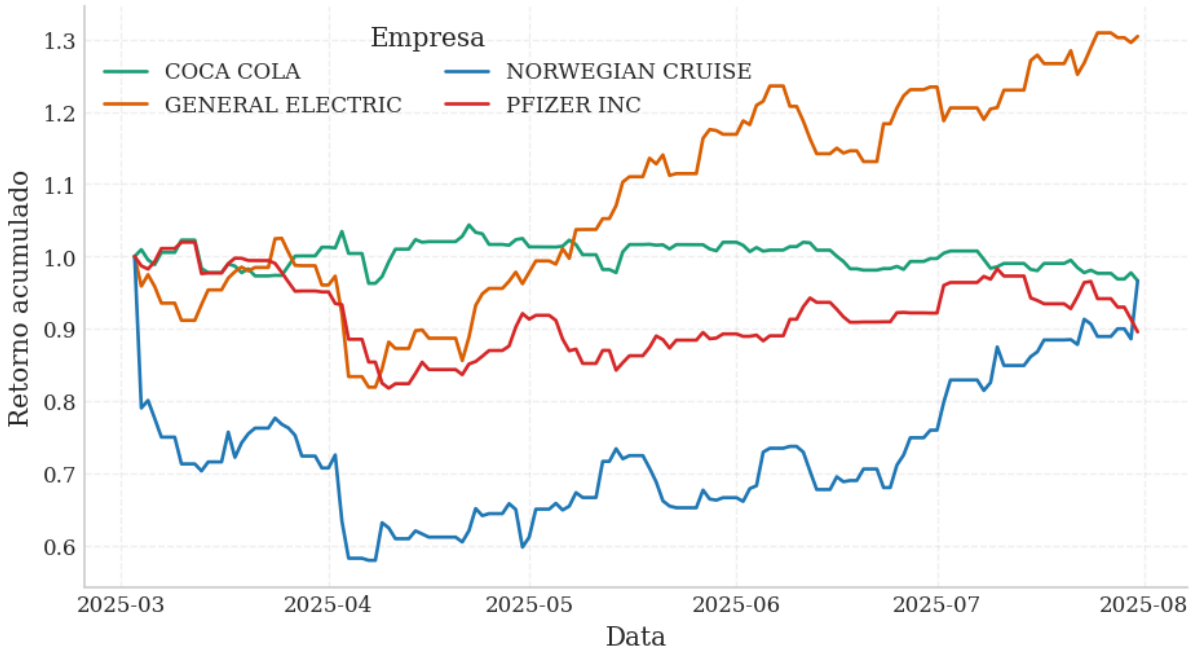


Figura IV.8: Gráfico ilustrando a evolução da valorização de empresas pertencentes a diferentes setores da economia no período de 03/03/2025 a 31/07/2025. A figura permite comparar o comportamento de mercados distintos ao longo do tempo, destacando divergências de desempenho, tendências específicas de cada setor e momentos de maior volatilidade relativa.

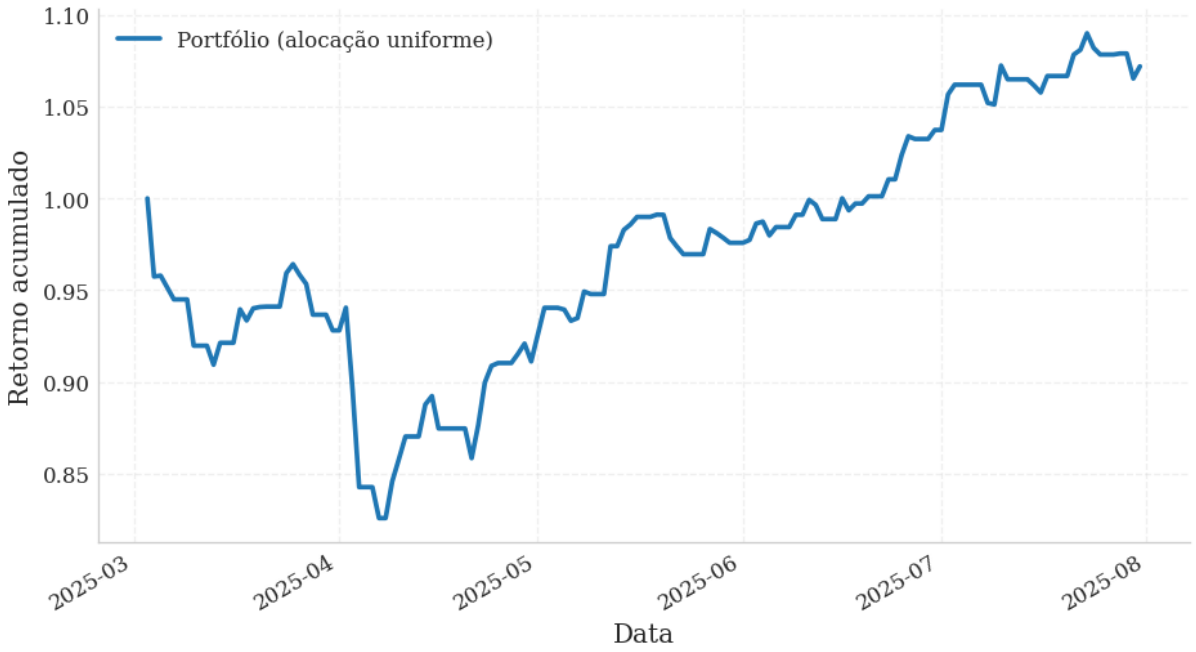


Figura IV.9: Gráfico apresentando a evolução da valorização de um portfólio balanceado composto pelas ações das 19 empresas analisadas neste estudo, ao longo do período de 03/03/2025 a 31/07/2025. A figura evidencia o comportamento agregado do portfólio, permitindo observar sua estabilidade, tendências de crescimento e possíveis variações resultantes da diversificação entre diferentes setores.

Tabela IV.5: Tendência dos retornos acumulados, avaliada por meio do teste de Mann-Kendall, acompanhado dos respectivos valores-p. O teste foi conduzido considerando um nível de significância $\alpha = 0{,}05$, permitindo verificar a presença de tendências monotônicas significativas nas séries analisadas.

Empresa	**Tendência (MK)**	**Valor-p**
American Airlines Group	Crescente	$3,30 \cdot 10^{-3}$
Apple	Sem tendência	$0,07$
Bank of America	Crescente	$0,0$
Cisco Systems	Crescente	$2,88 \cdot 10^{-15}$
Citigroup	Crescente	$0,0$
Coca-Cola	Decrescente	$6,23 \cdot 10^{-3}$
Delta Air Lines	Crescente	$8,67 \cdot 10^{-8}$
Ford Motor	Crescente	$0,0$
Freeport-McMoRan	Crescente	$8,43 \cdot 10^{-15}$
General Electric	Crescente	$0,0$
General Motors	Crescente	$8,87 \cdot 10^{-10}$
Intel	Sem tendência	$0,45$
JPMorgan Chase	Crescente	$0,0$
Microsoft	Crescente	$0,0$
Morgan Stanley	Crescente	$0,0$
Newmont	Crescente	$0,0$
Norwegian Cruise	Crescente	$7,61 \cdot 10^{-6}$
Pfizer	Sem tendência	$0,69$
Tesla	Crescente	$8,82 \cdot 10^{-11}$

Capítulo V Metodologia para Construção do Agente de Negociação Intradiária via Aprendizado por Reforço Profundo

Neste capítulo, apresentamos a metodologia proposta para resolver a principal questão abordada nesta dissertação, a saber, a adequação de algoritmos de ARP para problemas de negociação intradiária com múltiplas ações. A Seção V.1 apresenta a formulação do problema analisado. A Seção V.2 trata da modelagem do PDM necessário para o agente. Por fim, na Seção V.3, apresentamos os experimentos realizados para o treinamento e teste do agente, bem como as métricas utilizadas em sua avaliação.

V.1 Formulação do problema

Em consonância com a Seção I.2, o problema investigado nesta pesquisa consiste em avaliar se um algoritmo de Aprendizado por Reforço Profundo (ARP) é capaz de apresentar desempenho superior, no contexto de negociação intradiária envolvendo múltiplas ações, quando comparado a estratégias de referência amplamente utilizadas, como *buy and hold* e índices de mercado.

Sob a perspectiva do Aprendizado por Reforço (AR), esse problema pode ser estruturado em dois componentes interdependentes. O primeiro refere-se à capacidade do agente de estimar, a partir do estado observado, a probabilidade de variações futuras no preço de múltiplos ativos. Essa estimativa é essencial para a construção de uma função de valor ou de uma política parametrizada. O segundo componente diz respeito ao processo de tomada de decisão sequencial: dado um estado s_t, o agente seleciona uma ação

$$a_t \in [-1, 1], \quad \begin{cases} a_t > \delta & \text{indica ação de compra,} \\ a_t < -\delta & \text{indica ação de venda,} \\ -\delta \leq a_t \leq \delta & \text{indica manutenção da posição.} \end{cases}$$

A escolha da ação visa maximizar a recompensa acumulada, definida pela função $R(s_t, a_t)$. O agente interage com o ambiente ao longo de episódios e aprende uma política $\pi(a \mid s)$ que orienta suas ações com base na dinâmica de transição dos estados e nas recompensas observadas, buscando otimizar seu desempenho em um ambiente dinâmico e estocástico como o mercado

financeiro.

A integração entre esses dois componentes, previsão de variações futuras e tomada de decisão sequencial, define a formulação central do problema. O agente deve, simultaneamente, inferir a evolução provável dos preços e determinar a ação ótima entre comprar, vender ou manter sua posição. A Figura V.1 ilustra esse fluxo de decisão.

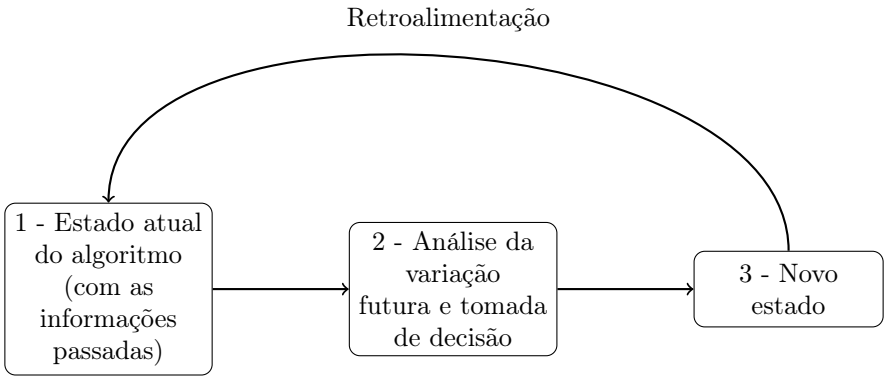


Figura V.1: Fluxo do processo de decisão do agente no PDM. Cada caixa numerada corresponde a uma etapa da dinâmica de interação: (1) o agente observa o estado atual, composto pelas informações históricas relevantes; (2) realiza a análise da variação futura, isto é, estima valores esperados para o próximo instante e avaliar as três ações disponíveis (comprar, manter ou vender); e (3) a partir da decisão selecionada, o ambiente retorna um novo estado, reiniciando o ciclo. A seta de retroalimentação ilustra a evolução temporal do processo, indicando que cada decisão atual influencia diretamente as próximas observações do agente.

Consideramos um ambiente no qual, no tempo t, o agente observa um conjunto de informações referentes aos preços e demais variáveis de múltiplos ativos financeiros. A partir dessas observações, o objetivo é que o agente aprenda uma política $\pi(a \mid s)$ capaz de orientar decisões no instante subsequente $t + 1$, buscando maximizar o retorno acumulado ao longo do tempo.

Durante o processo de aprendizado, o agente interage repetidamente com o ambiente, recebe recompensas em função das ações tomadas e atualiza seus parâmetros com o intuito de melhorar seu desempenho em episódios futuros. Essa formulação assume um cenário intrinsecamente dinâmico e estocástico, no qual a incerteza dos movimentos do mercado torna a decisão sequencial um elemento central do problema. Compete ao agente explorar e identificar oportunidades, ajustando sua política a fim de maximizar o retorno esperado.

Além disso, a formulação do problema contempla dois cenários experimentais distintos: ativo único e múltiplos ativos. No cenário de ativo único, o agente interage com o ambiente considerando apenas um único ativo financeiro. Nesse caso, a ação a_t representa exclusivamente a decisão de comprar, manter ou vender o ativo em questão, conforme a discretização definida pelo limiar δ. Esse cenário permite avaliar a capacidade do agente de identificar pontos de entrada e saída, explorando a dinâmica temporal dos preços de um único ativo e maximizando o retorno

associado à sequência de decisões tomadas.

No cenário de múltiplos ativos, o problema apresenta maior complexidade, pois o agente deve tomar decisões simultâneas envolvendo múltiplos ativos. Nesse contexto, a ação no instante t passa a ser representada por um vetor

$$\mathbf{a}_t = \left(a_t^{(1)}, a_t^{(2)}, \ldots, a_t^{(N)}\right),$$

em que N denota o número de ativos disponíveis e cada componente $a_t^{(i)} \in [-1, 1]$ indica a decisão associada ao ativo i, seguindo a mesma interpretação de compra, venda ou manutenção. Além disso, o módulo de cada componente está associado à intensidade da ação, refletindo a fração de capital a ser alocada ou desalocada em cada ativo.

Dessa forma, além de decidir entre comprar, vender ou manter posições individualmente, o agente deve aprender a balancear o portfólio ao longo do tempo, selecionando quais ativos devem receber maior ou menor alocação de recursos. Esse processo envolve a coordenação das decisões individuais em uma estratégia global de alocação, respeitando restrições de capital e buscando maximizar o retorno agregado do portfólio. O cenário múltiplos ativos, portanto, aproxima a formulação do problema de situações reais de gestão de portfólio, nas quais o desempenho depende não apenas da decisão sobre cada ativo isoladamente, mas também da interação entre eles.

A avaliação empírica e a comparação com estratégias de referência serão discutidas detalhadamente no Capítulo VI.

V.2 Especificação do Processo de Decisão de Markov

Para o desenvolvimento do agente de negociação intradiária, foi necessário formalizar a estrutura do Processo de decimal de Markov (PDM) que fundamenta sua interação com o ambiente. Essa modelagem envolve a definição do espaço de estados, do conjunto de ações, da função de recompensa, das condições iniciais e dos estados terminais, que em conjunto determinam o comportamento dinâmico do processo de aprendizado.

V.2.1 Estados

O estado s_t utilizado pelo agente é construído conforme detalhado na Seção IV.3. Cada observação x_t contém as variáveis provenientes do livro de ofertas, das últimas transações e dos dados internos do agente, conforme listadas na Tabela IV.3. Além disso, o estado incorpora uma janela temporal com μ observações consecutivas, formando o tensor final $s_t \in \mathbb{R}^{\mu \times d}$.

Assim, o espaço de estados é composto por uma representação histórico-estrutural da microestrutura do mercado, acrescida de informações sobre a posição e o patrimônio do agente. Essa representação fornece o contexto necessário para que a política aprenda padrões temporais e reaja às condições do ambiente.

Do ponto de vista operacional, a dinâmica dos estados segue o fluxo natural das transições temporais: a cada novo registro do ambiente, um novo vetor x_t é incorporado à janela deslizante, enquanto a observação mais antiga é descartada. Esse processo garante consistência temporal e mantém o estado atualizado com as variações recentes de preço, liquidez e posição.

Os estados terminais ocorrem em duas situações específicas:

- **Queda do patrimônio abaixo do limite mínimo**: quando a perda acumulada do agente ultrapassa o limiar ε, estipulado como fração do capital inicial.
- **Exaustão dos dados**: quando o agente alcança a última observação disponível no conjunto de dados intradiários.

Essas condições delimitam o horizonte do episódio e fornecem sinalizadores claros para o agente sobre o encerramento de sua trajetória.

V.2.2 Ações

O espaço de ações é contínuo no intervalo $[-1, 1]$. Cada ação a_t representa a intenção do agente em relação à sua posição no ativo:

- Intervalo $[-1, -\delta[$: aciona operações de venda;
- Intervalo $]\delta, 1]$: aciona operações de compra;
- Intervalo $[-\delta, \delta]$: nenhuma operação é executada.

Para evitar operações decorrentes de ruídos numéricos, utiliza-se um limiar δ que define zonas de inação. Assim, somente ações superiores a δ ou inferiores a $-\delta$ resultam em compras ou vendas, respectivamente. A magnitude da ação determina a fração do saldo (no caso de compras) ou da posição atual (no caso de vendas) a ser movimentada.

Esse formato contínuo justifica o uso de algoritmos apropriados para ambientes contínuos, tais como PPO, *Advantage Actor-Critic* (A2C), DDPG e *Soft Actor-Critic* (SAC), empregados nos experimentos deste trabalho.

V.2.3 Função Recompensa

Após definir o espaço de estados e ações, foi necessário selecionar uma função de recompensa que incentivasse o agente a adotar decisões lucrativas. Três estratégias principais foram analisadas:

- Função local;
- Função de lucro líquido
- Função diferencial de lucro líquido

V.2.3.1 Função Local

Avalia a recompensa com base nas transições entre estados consecutivos. O cálculo considera a quantidade negociada e a variação de preço entre as observações anterior e posterior à ação do agente. A Equação V.1 demonstra o cálculo. Nessa equação, r_t é a recompensa obtida, q é a quantidade de cotas transacionadas, p_{t-1} é o preço anterior e p_t é o preço subsequente à ação.

$$r_t = q \cdot (p_{t-1} - p_t) \tag{V.1}$$

V.2.3.2 Função de Lucro Líquido

Essa abordagem oferece uma visão agregada do desempenho financeiro do agente em uma transição específica, refletindo diretamente o lucro ou prejuízo acumulado no momento. A Equação V.2 demonstra a função. Nessa equação, l_{bt} representa o lucro bruto da transição atual, θ representa o valor inicial disponível para operação pelo agente, t_t representa o valor das taxas na transição atual.

$$l_t = l_{bt} - \theta - t_t \tag{V.2}$$

V.2.3.3 Função Diferencial de Lucro Líquido

Esta função avalia a variação no lucro líquido entre duas transições consecutivas. Diferentemente da função local, pode atribuir recompensas mesmo na ausência de negociações, com base na valorização ou desvalorização das cotas mantidas pelo agente. A Equação V.3 demonstra a função. Nessa equação, l_{t-1} é o lucro líquido da transição anterior e l_t da transição atual.

$$r_t = l_{t-1} - l_t \tag{V.3}$$

As seleções dessas funções de recompensa exerce impacto direto na estabilidade do treinamento e na qualidade da política aprendida. No contexto de negociação intradiária, optamos por métricas baseadas em lucro líquido e seu diferencial porque elas capturam não apenas o resultado imediato de uma transação, mas também o efeito acumulado das decisões do agente ao longo do episódio.

V.2.4 Parâmetros da Modelagem

Os principais parâmetros utilizados na definição do PDM encontram-se na Tabela V.1. Esses parâmetros incluem o capital inicial θ, o limiar de término por perda ε, a profundidade temporal μ da janela de observações e o limiar de sensibilidade da ação δ. Em conjunto, eles delimitam o comportamento operacional do ambiente e como o agente pode interagir ao longo de cada episódio.

Tabela V.1: Parâmetros adotados para a definição do PDM, abrangendo as variáveis estruturais e operacionais que determinam o comportamento do ambiente de decisão e orientam as interações do agente.

Notação	Descrição
α	Fator multiplicativo utilizado para ajustar a taxa de lucro líquido em relação ao montante investido, influenciando o cálculo das recompensas por transição e por episódio.
θ	Quantia inicial disponibilizada ao agente para execução de operações no ambiente de simulação.
ε	Porcentagem do valor inicial que, ao ser perdida, determina a condição de término de um episódio.
μ	Número de observações que compõem o estado atual do agente.
d	Profundidade da memória do agente, indicando a quantidade de observações anteriores incluídas no estado atual.
δ	Limiar para discretização da ação contínua, determinando a sensibilidade do agente ao decidir entre comprar, vender ou manter posição.

V.3 Treinamento e validação do agente

Tanto o processo de treinamento quanto o de avaliação do modelo utilizamos os algoritmos PPO, A2C, SAC e DDPG. A escolha por estes métodos de Aprendizado por Reforço Profundo se devem às suas compatibilidades com ambientes que apresentam espaços de observação e ação contínuos, como o agente modelado neste trabalho.

V.3.1 Configuração dos Hiperparâmetros do algoritmo

A implementação dos algoritmos (PPO, A2C, DDPG e SAC) utilizados neste trabalho segue a configuração oferecida pela biblioteca Stable Baselines3, com adaptações específicas ao problema de negociação intradiária apresentado. Os principais parâmetros empregados na construção e treinamento do agente são:

- `MlpPolicy`: Define que a política do agente é parametrizada por uma rede neural multicamadas (*Multilayer Perceptron*). Essa escolha é adequada para problemas de decisão com observações numéricas e contínuas, como os estados derivados dos dados de mercado coletados, evitando a necessidade de arquiteturas mais complexas, como redes convolucionais.

- `env_training`: Ambiente vetorizado utilizado pelo algoritmo, implementado por meio da classe `DummyVecEnv`. Mesmo quando apenas uma instância do ambiente é utilizada, o algoritmo exige que a interação ocorra por meio de um *vectorized environment*, pois seus mecanismos internos de coleta de transições, armazenamento no *rollout buffer* e atualização das políticas são estruturados para operar sobre lotes de observações e recompensas. O `DummyVecEnv` encapsula o ambiente do PDM e o adapta para o formato requerido pelo algoritmo, padronizando chamadas de `reset` e `step`, organizando observações, recompensas e sinais de término em tensores vetorizados. Dessa forma, cada interação do agente com o ambiente é registrada de forma consistente, garantindo compatibilidade com a arquitetura da política (`MlpPolicy`) e com o processo de otimização do algoritmo.

- `learning_rate = 1e-5`: Taxa de aprendizado utilizada pelo otimizador da rede neural. O valor escolhido (0,00001) é relativamente baixo, o que reduz a magnitude das atualizações de gradiente. Essa opção é especialmente relevante em ambientes financeiros, onde pequenas variações nos dados podem gerar descontinuidades na superfície de otimização. Uma taxa de aprendizado menor oferece maior estabilidade ao treinamento, evitando oscilações e divergência da política.

Em conjunto, esses parâmetros buscam garantir um treinamento estável do algoritmo, reduzindo oscilações no processo de maximização da vantagem e acomodando a sensibilidade das séries temporais de preços. A rede totalmente conectada escolhida se mostrou suficiente para capturar padrões intra-diários, enquanto o baixo *learning rate* e o monitoramento contínuo via TensorBoard ajudam a mitigar comportamentos indesejados durante o processo de otimização das políticas.

V.3.2 Validação Cruzada

Para conduzir o processo de treinamento e avaliação do agente, foi empregada uma estratégia de validação cruzada adequada a séries temporais. Essa escolha se justifica pelo fato de que, nesse tipo de problema, o conjunto de observações $(x_t, y_t)_{t=1}^{T}$ apresenta dependência ao longo do tempo: os valores em um instante t são influenciados, direta ou indiretamente, pelas informações registradas em momentos anteriores.

Para evitar o uso indevido de informações futuras, adotou-se a validação cruzada temporal, que preserva integralmente a sequência cronológica dos dados. Essa abordagem reproduz um cenário mais próximo da prática, no qual o modelo é ajustado inicialmente com um segmento histórico e, em seguida, é avaliado em períodos posteriores, respeitando a evolução natural da série.

No presente trabalho, a validação cruzada foi configurada com cinco dobras, nas quais o conjunto histórico de cada ativo foi particionado em cinco blocos temporais consecutivos, formando pares de treino e teste sem qualquer vazamento de informação futura. Em cada iteração, a separação entre os conjuntos segue a lógica

$$\underbrace{\text{Treino}}_{\text{tudo até } t_k} \quad \Big| \quad \underbrace{\text{Teste}}_{\text{intervalo imediatamente seguinte}}$$

de modo que o conjunto de treinamento se expande progressivamente a cada dobra, enquanto o conjunto de teste corresponde sempre ao período subsequente. Esse procedimento garante que, em todas as iterações, o agente treine exclusivamente com informações do passado e seja avaliado apenas sobre períodos posteriores, conforme exemplificado na Figura V.2.

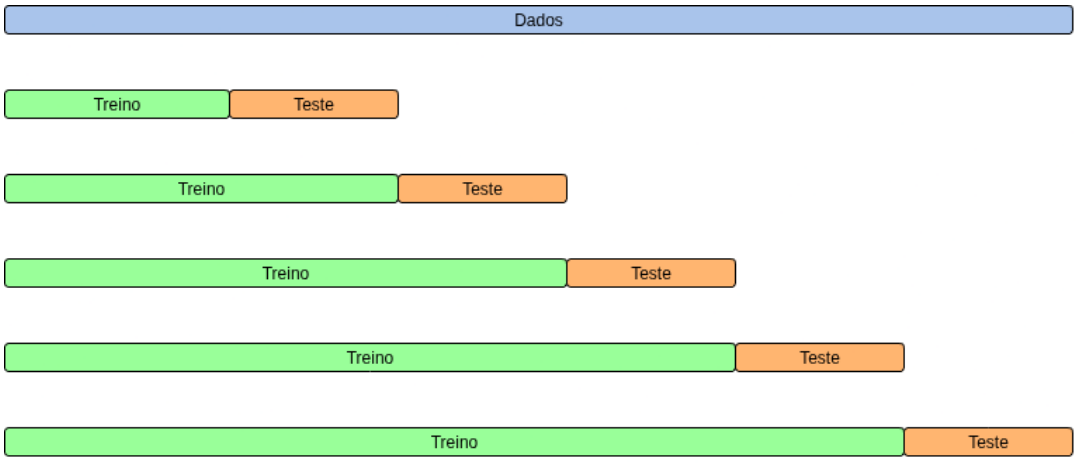


Figura V.2: Validação cruzada para séries temporais é uma estratégia de validação que preserva a ordem temporal, dividindo os dados em subconjuntos sequenciais para evitar vazamento de informações e avaliar o desempenho do modelo em múltiplos pontos no tempo.

A divisão é realizada programaticamente da seguinte forma: Cada dobra gera dois blocos distintos, um conjunto de treinamento, contendo observações até o instante t_k, e um conjunto de teste, correspondente ao período subsequente. Esses blocos são salvos temporariamente e utilizados nas etapas de treino e validação do agente, respectivamente. O agente é treinado com o conjunto de treino e, posteriormente, avaliado apenas sobre os dados do conjunto de teste, sem atualização adicional de parâmetros.

Dessa forma, cada experimento é composto por:

1. uma fase de aprendizado, em que o agente interage com o ambiente utilizando exclusivamente os dados do bloco de treinamento;

2. uma fase de teste, em que o agente age de maneira puramente avaliativa no bloco temporal imediatamente posterior;

3. a repetição desse processo para as cinco dobras

A Tabela V.2 detalha a divisão feita nos períodos e as proporções dos dados.

Tabela V.2: Esquema ilustrando a divisão das dobras na validação cruzada temporal. A tabela mostra como os conjuntos de treino e teste são organizados ao longo da linha do tempo. As datas de início e fim como as proporções dos dados são vistos na tabela

Dobra	**Início Treino**	**Fim Treino**	**Início Teste**	**Fim Teste**	**Treino**	**Teste**
1	2025-03-03	2025-03-26	2025-03-26	2025-04-22	50%	50%
2	2025-03-03	2025-04-22	2025-04-22	2025-05-12	66.6%	33.3%
3	2025-03-03	2025-05-12	2025-05-12	2025-06-05	75%	25%
4	2025-03-03	2025-06-05	2025-06-05	2025-07-03	80%	20%
5	2025-03-03	2025-07-03	2025-07-03	2025-07-31	83.3%	16.6%

V.3.3 Métricas utilizadas

Durante os ciclos de treinamento e teste, diferentes métricas são registradas e armazenadas, possibilitando uma análise aprofundada do desempenho do agente. Esses dados incluem tanto visualizações gráficas quanto arquivos CSV, que contêm os principais indicadores obtidos.

As métricas de desempenho extraídas durante os experimentos (ver Capítulo VI) são descritas a seguir.

1. **Média do lucro líquido do teste.** Esta métrica representa a média dos lucros líquidos observados durante o período de teste. O lucro líquido é calculado subtraindo-se os custos totais (como taxas de transação e custos operacionais) da receita obtida com as negociações

realizadas. A média do lucro líquido oferece uma visão geral do desempenho do agente em termos de rentabilidade ao longo de várias iterações ou episódios de teste.

2. **Desvio padrão do lucro líquido do teste.** O desvio padrão do lucro líquido mede a volatilidade ou a dispersão dos lucros líquidos obtidos ao longo do teste. Quanto maior o desvio padrão, maior a variabilidade nos resultados, o que pode indicar um comportamento de negociação mais arriscado ou imprevisível. Esse valor ajuda a entender não apenas a rentabilidade média do agente, mas também sua consistência ao longo do tempo.

3. **Máxima perda acumulada do teste.** A máxima perda acumulada é a maior queda observada no valor do portfólio durante o período de teste, calculada como a diferença entre o pico mais alto e o fundo mais baixo antes de uma recuperação. Esta métrica é crucial para avaliar o risco do agente, pois ela indica a maior perda que o agente pode sofrer antes de começar a recuperar o valor perdido. Uma perda acumulada maior sugere que o agente pode estar se expondo a riscos excessivos.

4. **Retorno do teste.** O retorno do teste é a variação percentual entre o valor final e o valor inicial do portfólio durante o período de teste. Este valor fornece uma visão geral do desempenho total do agente, refletindo o quanto ele foi capaz de aumentar ou diminuir seu capital ao longo do tempo. É uma métrica chave para avaliar a eficácia do agente.

5. **Retorno dos índices comparativos.** Esta métrica compara o desempenho do agente com o retorno de índices de mercado amplamente reconhecidos. O retorno dos índices comparativos serve como uma referência para avaliar se a estratégia do agente é superior ou inferior aos benchmarks do mercado financeiro. O desempenho do agente é analisado em relação a esses índices para determinar sua eficácia relativa em condições de mercado comuns.

Além das métricas tradicionais de desempenho, torna-se essencial avaliar o agente também sob a ótica risco-retorno. Em operações de negociação intradiária, a volatilidade intradiária pode afetar significativamente o desempenho de qualquer estratégia. Assim, métricas que capturam apenas o lucro acumulado não são suficientes para refletir a qualidade de uma política aprendida.

A análise risco-retorno permite compreender se a estratégia é capaz de gerar resultados consistentes diante da variabilidade do mercado. Como a perda acumulada e o desvio padrão do lucro evidenciam a exposição ao risco, enquanto o retorno total sintetiza o desempenho acumulado, essas métricas complementares fornecem um panorama mais robusto do comportamento do agente.

Dessa forma, a adoção conjunta de retorno e medidas de risco é fundamental para comparar o agente de Aprendizado por Reforço com estratégias clássicas de mercado, garantindo uma avaliação mais alinhada às práticas consolidadas em finanças quantitativas.

Capítulo VI Avaliação Experimental

Neste capítulo apresentamos a análise experimental detalhada do método proposto. Cada seção deste capítulo aborda aspectos específicos da avaliação experimental, desde a configuração do ambiente computacional até comparações com os métodos existentes. A Seção VI.1, são apresentadas as configurações principais adotadas para a implementação do agente de Aprendizado por Reforço. Essa seção inclui detalhes técnicos sobre a arquitetura do agente, os parâmetros utilizados durante o treinamento, e as configurações utilizadas no projeto. A Seção VI.2 apresenta uma análise específica da influência do fator de desconto γ no comportamento e no desempenho do agente. Nessa seção, são conduzidos experimentos comparativos com diferentes valores de γ, permitindo avaliar como a ponderação entre recompensas imediatas e futuras afeta a política aprendida, a estabilidade das decisões e o retorno acumulado ao longo do tempo. A Seção VI.3 é dedicada aos experimentos de sensibilidade às variáveis do estado, cujo objetivo é identificar quais informações provenientes dos dados de ofertas são mais relevantes para o desempenho do agente. Como nem todas as variáveis contribuem de forma equivalente ao processo de decisão, essa análise permite distinguir atributos essenciais de elementos redundantes, avaliando o impacto individual e coletivo de cada componente do estado. Os resultados obtidos auxiliam na compreensão da sensibilidade do agente às diferentes fontes de informação e orientam a seleção de variáveis mais eficientes para compor a representação final do ambiente. A Seção VI.4 apresenta uma avaliação de arquiteturas de Aprendizado por Reforço. Os desempenhos das arquiteturas são analisadas com base em métricas específicas, possibilitando verificar a que possui maior efetividade para o problema proposto. A Seção VI.5 apresenta uma avaliação dos resultados obtidos a partir dos experimentos conduzidos. Os desempenhos do agente são analisados com base em métricas específicas, possibilitando a verificação da efetividade do modelo proposto. Também são discutidos os principais achados, limitações observadas e possíveis direções para trabalhos futuros.

VI.1 Configuração de Hardware e Software

Os experimentos foram executados em uma máquina com arquitetura x86_64, processador AMD Ryzen 7 3700U com Radeon Vega Mobile Gfx 8 4-Core 2.3GHz, 12GB de RAM, placa de

vídeo AMD Radeon Vega 10 Graphics e sistema operacional Ubuntu 24.04.2 LTS.

O desenvolvimento de treinamento e testes levou em consideração bibliotecas e frameworks da linguagem Python. Optou-se pela implementação do programa de coleta e do treinamento do agente nessa linguagem, em razão da ampla disponibilidade de bibliotecas e frameworks, o que simplifica o processo de implementação. Nessa implementação, usamos os pacotes detalhados na Tabela VI.1 e na Tabela VI.2.

Tabela VI.1: Tabela de softwares, frameworks e bibliotecas utilizados na construção do scrapper de captura dos dados

Pacote	**Versão**	**Descrição**
Python	v3.10	Linguagem utilizada no projeto
Pandas	v2.1.4	Ferramenta de manipulação e análise de dados tabulares.
NumPy	v1.26.4	Ferramenta de operações numéricas no Python. Fornece arrays eficientes e operações de álgebra linear.
Pytz	v2023.3.post1	Responsável por tratar fusos horários com precisão, útil ao lidar com datas e horas em aplicações internacionais.
BeautifulSoup4	v.4.12.2	Responsável por extrair dados de HTML e XML.
Lxml	v.5.2.2	Alternativa rápida e robusta para parsing de HTML e XLM. Usado em conjunto com o BeatifulSoup.
Selenium	v.4.16.0	Usado para automação de acesso ao navegador para o scraping
Geckodrive-autoinstaller	v.0.1.0	Garante que o driver do navegador esteja instalado automaticamente, facilitando o uso do Selenium, sem necessidade de configuração manual.
SQLAlchemy	v.2.0.23	Ferramenta para interação com bancos relacionais usando ORM ou SQL puro.
PyMySQL	v.1.1.0	Conector leve e escrito em Python para se conectar a bancos MySQL/MariaDB.
Fastparquet	v.2023.10.1	Biblioteca para ler e gravar arquivos parquet.

VI.2 Experimentos de sensibilidade ao fator de desconto

Os experimentos de sensibilidade ao fator de desconto γ visam analisar como a preferência temporal do agente influência seu desempenho no contexto de negociação intradiária. Em Aprendizado por Reforço, o fator de desconto controla a importância relativa atribuída às recompensas futuras no cálculo do retorno esperado, desempenhando papel central no equilíbrio entre decisões de curto e longo prazo.

O procedimento experimental consistiu em manter inalteradas as demais configurações do ambiente e do algoritmo, variando exclusivamente o valor de γ. Foram avaliados três valores representativos: $\gamma = 0{,}9$, $\gamma = 0{,}95$ e $\gamma = 0{,}99$. Esses valores refletem diferentes horizontes

Tabela VI.2: Software, frameworks e bibliotecas utilizadas para construção do agente de AR

Pacote	**Versão**	**Descrição**
Python	v3.12	Linguagem utilizada no projeto.
Pandas	v2.2.3	Ferramenta de manipulação e análise de dados tabulares.
Matplotlib	v3.9.2	Ferramenta para geração de gráficos e visualizações de dados.
Statsmodels	v0.14.4	Biblioteca para modelagem estatística, testes e análise de séries temporais.
Stable-Baselines3	v2.3.2	Conjunto de implementações de algoritmos de Aprendizado por Reforço baseados em PyTorch.
PyMannKendall	v1.4.3	Implementa testes de tendência em séries temporais com base no teste de Mann-Kendall.
Gym	v0.26.2	Kit de desenvolvimento para criação e simulação de ambientes de Aprendizado por Reforço.
SB3-Contrib	v2.3.0	Extensões e contribuições adicionais para o Stable-Baselines3, incluindo algoritmos experimentais.
Shimmy	v0.2.1	Adaptador de ambientes compatíveis com Gym para outras APIs de simulação.
TensorBoard	v2.18.0	Ferramenta de visualização para métricas e processos de treinamento em aprendizado profundo.
Scikit-learn	v1.6.1	Biblioteca para implementação de algoritmos de aprendizado de máquina supervisionado e não supervisionado.
YFinance	v0.2.54	Biblioteca para obtenção de dados financeiros históricos e atuais da Yahoo Finance.
Lxml	v5.3.1	Ferramenta rápida para parsing de HTML e XML

temporais de decisão, desde uma estratégia mais míope, focada em retornos imediatos, até uma estratégia orientada ao desempenho acumulado ao longo do episódio.

A Tabela VI.3 apresenta um resumo dos resultados obtidos para cada configuração de γ, considerando métricas de retorno final e máxima perda acumulada.

Tabela VI.3: Resumo dos resultados dos experimentos de sensibilidade ao fator de desconto γ

Valor de γ	**Interpretação**	**Retorno (%)**
$\gamma = 0{,}9$	Ênfase em recompensas imediatas, horizonte de decisão curto	16,1%
$\gamma = 0{,}95$	Compromisso entre curto e médio prazo	16,0%
$\gamma = 0{,}99$	Forte valorização de recompensas futuras, horizonte longo	16,3%

A análise dos resultados evidencia diferenças no comportamento do agente em função do valor de γ. Quando configurado com $\gamma = 0{,}9$, o agente demonstrou um comportamento mais reativo e orientado ao curto prazo, priorizando recompensas imediatas. Embora essa configuração possa

resultar em ganhos pontuais, ela não se mostra a melhor estratégia a longo prazo, perdendo valorizações que as ações teriam ao longo do tempo.

Com $\gamma = 0{,}95$, o agente apresentou o pior desempenho entre as configurações avaliadas. Esse resultado pode ser explicado pelo fato de que esse valor intermediário do fator de desconto induz um comportamento híbrido, no qual o agente não explora de forma eficiente nem oportunidades de curto prazo, nem benefícios de longo prazo. Ao atribuir peso moderado às recompensas futuras, o agente tornou-se menos reativo a variações imediatas do mercado, reduzindo sua capacidade de capturar movimentos intradiários favoráveis. Ao mesmo tempo, esse nível de desconto não foi suficiente para compensar essa perda de reatividade com ganhos futuros consistentes, resultando em decisões que não maximizaram o retorno acumulado.

Por fim, a configuração com $\gamma = 0{,}99$ produziu o comportamento mais alinhado ao objetivo de maximização do desempenho acumulado. Nessa configuração, o agente demonstrou maior tolerância a variações temporárias negativas, aceitando perdas de curto prazo em troca de ganhos futuros mais consistentes. O maior retorno final observado indica que a valorização de recompensas futuras contribui para decisões mais estáveis e coerentes com a dinâmica do mercado intradiário considerado.

Em síntese, os experimentos de sensibilidade ao fator de desconto demonstram que a escolha de γ exerce influência direta sobre o perfil decisório e o desempenho do agente. Valores mais baixos de γ incentivam estratégias excessivamente reativas, capazes de capturar ganhos imediatos, mas que tendem a sacrificar valor ao longo do tempo ao ignorar a continuidade dos movimentos de mercado. Por outro lado, valores intermediários podem induzir comportamentos inconsistentes, nos quais o agente não explora de forma eficaz nem oportunidades de curto prazo, nem benefícios de longo prazo, resultando em desempenho inferior. Nesse contexto, a configuração com $\gamma = 0{,}99$ mostrou-se a mais adequada, ao favorecer decisões orientadas à maximização do retorno acumulado, com maior estabilidade e coerência frente à dinâmica intradiária do mercado. Esses resultados reforçam a importância de calibrar cuidadosamente o fator de desconto conforme o horizonte temporal e as características do problema de negociação considerado.

VI.3 Experimentos de sensibilidade às variáveis do estado

Os experimentos de sensibilidade às variáveis do estado constitui uma técnica usada para determinar o impacto de características específicas no desempenho do modelo em AM. Trata-se de um método que permite compreender o impacto real de cada informação sobre o desempenho preditivo do modelo, indo além de uma simples análise de importância global.

O procedimento consiste na remoção ou modificação sistemática de uma variável do conjunto

de dados, ou do processo de inferência, seguida pela reavaliação das métricas de desempenho. A diferença observada no desempenho permite inferir a relevância efetiva daquela informação para o modelo.

Tabela VI.4: Resumo dos resultados dos experimentos de sensibilidade às variáveis do estado das informações utilizadas no agente de ARP

Nome da informação	**Explicação da informação**	**Retorno (%)**	**Máx. Perda Acumulada**
Todas as informações (Baseline)	Todas as informações disponíveis	15,9%	-2,02
Removendo a coluna *Shares*	Quantitativo de ações negociadas	19,5%	-1,74
Removendo a coluna *Prices*	Valor das ações negociadas	3,8%	-3,19
Removendo as colunas *Time*	Informações de Data e Tempo	-19,2%	-3,17
Removendo a coluna *Last_10_Prices*	Preços das ações nas últimas 10 transações	-12,6%	-2,84
Removendo a coluna *Last_10_Shares*	Quantitativo das ações as últimas 10 transações	27,5%	-2,39

A remoção das diferentes informações revelou sinais claros sobre o que o agente estava realmente utilizando para tomar decisões. Quando removemos informações explícitas de preços, tanto a coluna *Prices* quanto o histórico *Last_10_Prices*, o desempenho caiu de forma significativa: o retorno e a máxima perda acumulada aumentou. Esse resultado indica que o agente depende fortemente de sinais diretos de preço e de sua série temporal recente para estimar tendências e reagir a movimentos de mercado. Sem esse contexto, o agente perde a capacidade de identificar entradas e saídas rentáveis e fica mais exposto a perdas excessivas, explicando a piora observada.

Algumas informações ligadas a posição/quantidade mostram efeito ambíguo. Ao retirar *Shares* (quantidade de ações), o agente melhorou ligeiramente, o que indica que essa variável talvez tenha introduzido ruído ou redundância com outras informações (por exemplo, sinais derivados do preço). Já a remoção de *Last_10_Shares* também aumentou o lucro em relação ao baseline, o que reforça a hipótese de que históricos de quantidade, no formato atual, não estavam acrescentando previsibilidade, possivelmente porque o volume ou as posições não eram informativos no horizonte temporal do agente, ou porque estavam correlacionados com outras entradas de forma que confundiam a política.

Finalmente, a remoção de *Time* teve efeito negativo, provavelmente porque com essa informação o agente seja capaz de capturar pequenas variações ou sincronizações do mercado (por exemplo, latência ou padrões em curtíssimo prazo) que o agente aproveita.

VI.4 Comparação de arquiteturas de Aprendizado por Reforço

Esta seção apresenta uma análise comparativa entre diferentes algoritmos de aprendizado por reforço aplicados ao contexto do mercado financeiro. O objetivo é avaliar o desempenho, as vantagens e as limitações de cada abordagem frente às características específicas do ambiente de negociação desenvolvido. São considerados algoritmos representativos de diferentes famílias de métodos, como o DDPG, o PPO e o SAC e o A2C de modo a compreender como distintas estratégias de exploração e atualização de políticas impactam os resultados obtidos. A partir dessa comparação, busca-se identificar quais arquiteturas apresentam maior estabilidade de aprendizado, eficiência no uso das recompensas e capacidade de generalização em cenários de alta variabilidade. Por fim, os resultados discutidos ao longo desta seção oferecem subsídios para a escolha do algoritmo mais adequado às características do problema e fundamentam as análises realizadas nos capítulos seguintes.

O algoritmo A2C é uma técnica de aprendizado por reforço que combina elementos dos métodos de política (ator) e de valor (crítico) para otimizar o processo de tomada de decisão de um agente. O ator é responsável por aprender uma política que define a probabilidade de selecionar cada ação em um determinado estado, enquanto o crítico avalia a qualidade dessa ação estimando o valor esperado do retorno futuro. A principal característica do A2C é o uso da vantagem, uma medida que indica o quanto uma ação é melhor do que a média esperada, reduzindo a variância do gradiente e acelerando a convergência do treinamento. Além disso, ele realiza atualizações de forma síncrona [Mnih et al., 2016].

O algoritmo DDPG é um método de aprendizado por reforço profundo desenvolvido para lidar com ambientes contínuos, nos quais as ações não são discretas, mas assumem valores reais. Ele combina ideias do *Deterministic Policy Gradient* (DPG) com a estrutura das DQN utilizando redes neurais profundas tanto para representar a política (ator) quanto a função de valor (crítico). O ator aprende uma política determinística que mapeia diretamente estados para ações, enquanto o crítico avalia a qualidade dessas ações estimando o retorno esperado. Para garantir estabilidade no treinamento, o DDPG emprega redes-alvo e um *replay buffer*, que armazena experiências passadas e permite o aprendizado por amostragem, reduzindo a correlação temporal dos dados. Além disso, ele introduz ruído nas ações para promover a exploração do ambiente, compensando a natureza determinística da política [Lillicrap et al., 2016].

O algoritmo SAC é um método de aprendizado por reforço profundo baseado em políticas *off-policy* que busca equilibrar *exploration* e *exploitation* de forma mais eficiente ao incorporar o conceito de entropia na função de recompensa. Diferentemente de algoritmos tradicionais que visam apenas maximizar o retorno esperado, o SAC adiciona um termo de entropia que incentiva o agente a manter suas ações mais estocásticas, promovendo uma exploração mais ampla do espaço de estados e evitando a convergência prematura para políticas subótimas. Ele utiliza três redes neurais principais: duas para o crítico, a fim de reduzir o viés de superestimação dos valores de ação, e uma para o ator, que aprende uma política estocástica parametrizada. Além disso, faz uso de redes-alvo e de um *replay buffer* para garantir maior estabilidade durante o aprendizado.

Tabela VI.5: Resultados comparativos entre arquiteturas de Aprendizado por Reforço

Algoritmo	**Lucro líquido do portfólio**	**Desvio Padrão do lucro líquido do portfólio**	**Retorno do portfólio**	**Máx. Perda Acumulada**
A2C	162,30	±9,93	13,83%	-1,57
SAC	224,90	±11,81	19,16%	-1,66
DDPG	222,890	±15,42	18,99%	-1,51
PPO	229,89	±13,16	**19,59%**	-1,37

Com base nos resultados obtidos e apresentados na Tabela VI.5, observa-se que o algoritmo PPO apresentou o melhor desempenho geral entre as arquiteturas avaliadas, tanto em termos de retorno do portfólio quanto de estabilidade do aprendizado. O PPO obteve o maior retorno percentual (19,59%) e a menor perda acumulada (-1,37), demonstrando maior eficiência na maximização dos lucros e na mitigação de perdas em comparação aos demais métodos.

Esses resultados indicam que a estratégia de atualização proximal empregada pelo PPO, que limita a magnitude das mudanças na política entre iterações, contribui para um aprendizado mais estável e menos suscetível a oscilações abruptas, característica essencial em cenários de alta volatilidade como a negociação intradiária. Além disso, por equilibrar adequadamente *exploration* e *exploitation*, o PPO mostra-se capaz de generalizar melhor a partir de dados de mercado ruidosos, ajustando-se de forma mais robusta às variações dinâmicas dos preços. Assim, conclui-se que, entre os algoritmos analisados, o PPO foi o que obteve o melhor desempenho, sendo a arquitetura utilizada nos próximos testes.

VI.5 Resultados

Esta seção tem como finalidade apresentar e analisar, de forma detalhada, os resultados obtidos a partir dos experimentos propostos. Durante o processo de treinamento e avaliação do agente, utilizando-se os dados descritos na Seção IV.4, foram coletadas e monitoradas diversas

métricas relevantes para compreender o desempenho do modelo. Entre elas, destacam-se:

- análise de ações do agente treinadas e testadas para uma ação específica;
- análise de ações do agente treinadas para uma ações e testadas em outra;
- análise de ações do agente para um portfólio de ações

VI.5.1 Treinamento e teste para uma mesma ação

O objetivo desta seção é avaliar se o agente, treinado e testado para uma mesma ação, consegue ter um desempenho superior ao método *buy and hold*.

O *buy and hold* é uma estratégia de investimento em que o investidor adquire ativos financeiros e os mantém em carteira por um longo período, independentemente das oscilações de curto prazo do mercado. Essa abordagem baseia-se na premissa de que, ao longo do tempo, os mercados tendem a se valorizar, permitindo que o investidor capture ganhos provenientes da valorização dos ativos e dos dividendos pagos. Ao evitar operações frequentes de compra e venda, o investidor reduz custos relacionados a taxas, impostos e erros de timing, que podem impactar negativamente o retorno final.

A intenção é comparar o agente treinado focado em negociação intradiária com um método comum do mercado financeiro, entendendo se no período analisado o agente possui um retorno maior que a métrica. Para exemplificar o funcionamento do algoritmo ao longo das iterações, a Figura VI.1 demonstra a evolução do preço do ativo e do lucro obtido nas transições feitas de negociação intradiária para a ação da Coca-Cola.

A Tabela VI.6 apresenta uma comparação abrangente entre o desempenho de um agente de AR e a estratégia *Buy and Hold* para um conjunto diversificado de empresas de diferentes setores econômicos. A análise enfatiza o valor final médio obtido, o lucro médio, o retorno percentual e o desvio padrão das operações para ambos os métodos, além da diferença percentual em pontos percentuais entre as duas abordagens.

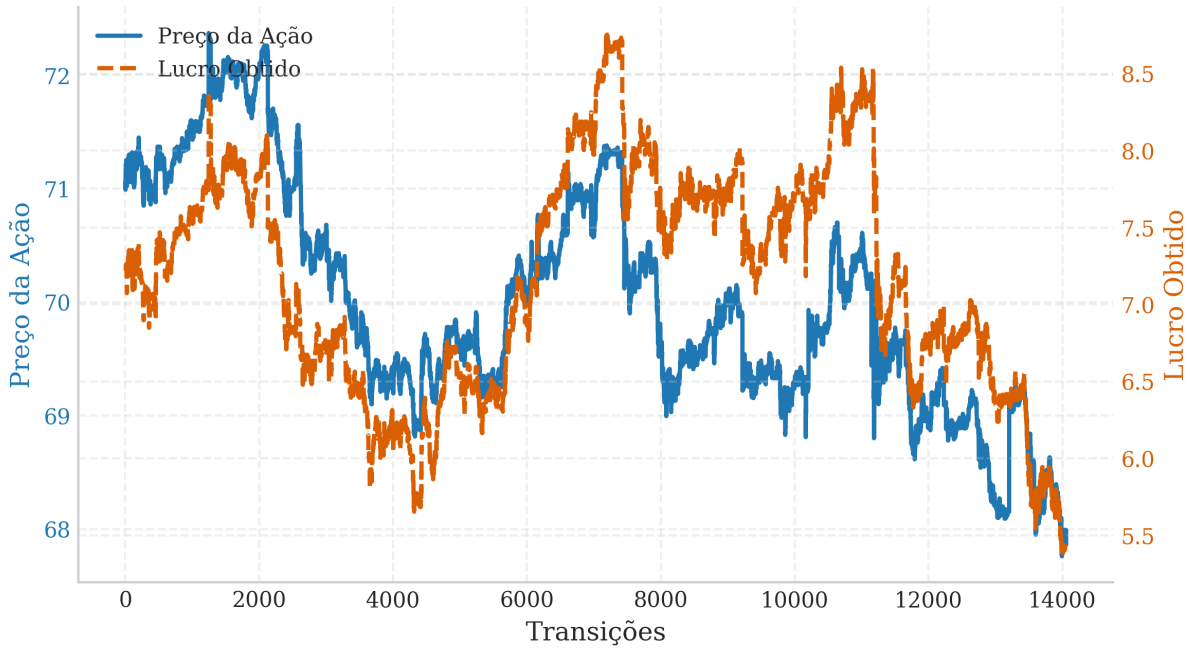


Figura VI.1: Gráfico ilustrando, ao longo das iterações do processo de análise, a variação do preço do ativo e o lucro (ou prejuízo) obtido pelo agente em cada transição. A figura permite visualizar como as decisões do agente se relacionam com a dinâmica de preços, evidenciando momentos de maior ganho, perda ou estabilidade durante a execução da estratégia.

Tabela VI.6: Resultados comparativos diários entre o agente de aprendizado por reforço e a estratégia *Buy and Hold* por empresa.

Empresa	**Valor Inicial**	**Valor Final Médio (Agente)**	**Média Lucro Agente**	**Retorno Agente (%)**	**Desvio Padrão Agente**	**% Iterações com Lucro**	**Média Lucro Buy & Hold**	**Retorno Buy & Hold (%)**	**Diferença (p.p)**
American Airlines	14,41	15,73	1,32	**9,13%**	0,91	90%	0,04	0,29%	**8,84**
Apple	241,65	258,96	17,31	**7,16%**	11,68	90%	-0,31	-0,13%	**7,29**
Bank of America	46,14	50,50	4,36	**9,46%**	0,43	100%	0,58	1,26%	**8,20**
Cisco	51,50	58,67	7,17	**13,92%**	1,82	100%	1,94	3,76%	**10,17**
Citigroup	49,78	57,54	7,76	**15,59%**	4,23	90%	4,94	9,93%	**5,65**
Coca-Cola	70,64	77,63	6,99	**9,90%**	1,19	100%	-0,16	-0,23%	**10,13**
Delta Air Lines	50,15	56,43	6,28	**12,51%**	3,13	100%	1,38	2,75%	**9,76**
Ford	9,63	10,74	1,11	**11,52%**	0,45	100%	0,14	1,50%	**10,02**
Freport-Mcmoran	55,00	58,98	3,98	**7,24%**	3,88	90%	0,08	0,15%	**7,09**
General Motors	42,00	47,60	5,60	**13,34%**	1,72	100%	0,93	2,22%	**11,12**
Intel Corp	24,50	26,37	1,87	**7,64%**	1,25	80%	-0,59	-2,39%	**10,03**
JP Morgan	256,03	286,58	30,55	**11,93%**	9,82	100%	6,47	2,53%	**9,41**
Microsoft	398,91	444,90	45,99	**11,53%**	13,32	100%	15,14	3,80%	**7,73**
Morgan Stanley	141,97	155,50	13,53	**9,53%**	6,74	90%	2,61	1,84%	**7,69**
Newmont	43,25	49,71	6,46	**14,93%**	2,88	100%	3,28	7,59%	**7,34**
Norwegian	26,57	28,51	1,94	**7,29%**	0,18	100%	0,80	3,02%	**4,27**
Pfizer	26,32	28,42	2,10	**7,98%**	1,20	90%	-0,23	-0,88%	**8,86**
Tesla	303,88	324,60	20,72	**6,82%**	13,72	80%	6,93	2,28%	**4,54**

Tabela VI.7: Resultados dos testes estatísticos comparando o desempenho do agente de aprendizado por reforço com a estratégia *Buy and Hold* por empresa.

Empresa	**Teste Estatístico**	**Média Agente**	**Média Buy & Hold**	**p-valor**	**Teste Significativo?**	**IC95 Agente**	**IC95 Buy & Hold**
American Airlines	t de Student	1,32	0,04	0,006	Sim	(0,75; 1,88)	(-0,65; 0,73)
Apple	Mann–Whitney	17,31	-0,31	0,002	Sim	(10,07; 24,55)	(-9,72; 9,10)
Bank of America	Mann–Whitney	4,36	0,58	0,001	Sim	(4,10; 4,63)	(-1,47; 2,63)
Cisco	Mann–Whitney	7,17	1,94	0,004	Sim	(6,04; 8,30)	(-0,91; 4,78)
Citigroup	t de Student	7,76	4,94	0,192	Não	(5,14; 10,38)	(-0,58; 10,47)
Coca-Cola	Mann–Whitney	6,99	-0,16	0,000	Sim	(6,26; 7,73)	(-0,92; 0,59)
Delta Air Lines	Mann–Whitney	6,28	1,38	0,027	Sim	(4,34; 8,21)	(-2,36; 5,12)
Ford	Mann–Whitney	1,11	0,14	0,001	Sim	(0,83; 1,39)	(-0,24; 0,53)
Freeport-Mcmoran	t de Student	3,98	0,08	0,085	Não	(1,58; 6,39)	(-4,60; 4,77)
General Motors	t de Student	5,60	0,93	0,012	Sim	(4,54; 6,67)	(-1,20; 3,07)
Intel	Mann–Whitney	1,87	-0,59	0,003	Sim	(1,09; 2,65)	(-1,88; 0,71)
JP Morgan	Mann–Whitney	30,55	6,47	0,001	Sim	(24,46; 36,64)	(-2,70; 15,63)
Microsoft	t de Student	45,99	15,14	0,000	Sim	(37,74; 54,25)	(4,13; 26,15)
Morgan Stanley	t de Student	13,53	2,61	0,002	Sim	(9,35; 17,70)	(-2,08; 7,30)
Newmont	t de Student	6,46	3,28	0,037	Sim	(4,67; 8,24)	(0,56; 6,00)
Norwegian	t de Student	1,94	0,80	0,062	Não	(1,83; 2,05)	(-0,50; 2,11)
Pfizer	Mann–Whitney	2,10	-0,23	0,001	Sim	(1,36; 2,84)	(-1,14; 0,68)
Tesla	Mann–Whitney	20,72	6,93	0,038	Sim	(12,22; 29,23)	(-11,50; 25,35)

A partir dos resultados apresentados, observa-se que o agente de aprendizado por reforço superou a estratégia *Buy and Hold* em todas as empresas avaliadas. Em diversas situações, o agente demonstrou capacidade de adaptar suas decisões de compra e venda a diferentes condições de mercado, resultando em retornos substancialmente superiores. Destacam-se, por exemplo, General Motors e Intel, que apresentaram diferenças positivas de 11,12 p.p. e 10,03 p.p., respectivamente, quando comparadas ao retorno da estratégia tradicional.

Além disso, o agente exibiu uma alta consistência operacional, atingindo lucro em até 100% das iterações para a maioria das empresas. Mesmo nos casos menos favoráveis, como Intel e Tesla, o agente manteve desempenho positivo em 80% das iterações, o que evidencia robustez nas decisões tomadas ao longo dos episódios. O comportamento do agente também se destaca na gestão de risco: na maioria das empresas, a máxima perda acumulada permaneceu em níveis reduzidos, indicando menor exposição a perdas durante a trajetória de decisão.

Além da análise de desempenho bruto, os resultados estatísticos reforçam de maneira robusta a superioridade do agente em relação à estratégia *Buy and Hold*. Para cada empresa, foram aplicados os testes t de Student e Mann-Whitney para comparar os retornos médios das duas abordagens. A hipótese nula ((H_0) considera que não há diferença entre os retornos médios, isto é, $H_0 : \mu_{\text{agente}} \leq \mu_{\text{Buy and Hold}}$, enquanto a hipótese alternativa (H_1) pressupõe que o retorno médio do agente seja superior ao da estratégia tradicional, $H_1 : \mu_{\text{agente}} > \mu_{\text{Buy and Hold}}$. Os resultados mostraram diferenças estatisticamente significativas para a maioria das empresas, com p-valor frequentemente inferiores a 0,01. Empresas como Apple, Microsoft, JPMorgan e Coca-Cola apresentaram efeitos altamente significativos, indicando que a vantagem do agente não se deve ao acaso, mas a um comportamento sistemático de melhor tomada de decisão. Em poucos casos, como Citigroup, Freeport-McMoRan e Norwegian, não foi possível rejeitar a hipótese nula ao nível de 5%, embora o agente ainda apresentasse desempenho superior em termos de retorno médio. Essa combinação de elevada significância estatística, intervalos de confiança favoráveis e consistência prática consolida a evidência de que o agente de AR oferece ganhos tanto economicamente relevantes quanto estatisticamente sustentados.

Em síntese, os resultados indicam que o agente de AR oferece uma abordagem adaptativa, com bom gerenciamento de risco e potencialmente mais lucrativa para negociação de ações. A capacidade de maximizar retornos, limitar quedas expressivas e gerar lucro de forma consistente na maioria dos testes reforça o potencial dessa metodologia como alternativa superior à estratégia tradicional de *Buy and Hold* na maior parte dos cenários analisados.

VI.5.2 Treinamento e teste para ações diferentes

O objetivo desta seção é analisar a capacidade de generalização do agente, ou seja, verificar se um modelo treinado em determinada ação consegue apresentar bom desempenho quando aplicado a uma ação diferente. Essa abordagem busca avaliar se o aprendizado adquirido pelo agente é específico de um único ativo ou se consegue capturar padrões mais amplos do mercado, tornando-se aplicável a diferentes contextos.

Dessa forma, o experimento consiste em treinar o agente utilizando os dados históricos de uma empresa e, posteriormente, testar seu desempenho em outra. Essa metodologia é relevante, pois reflete situações reais em que um modelo pode ser exposto a ativos distintos daqueles utilizados no treinamento, seja por mudanças de portfólio ou pela necessidade de avaliar diferentes setores do mercado.

Tabela VI.8: Resultados comparativos entre o agente de AR sendo treinado e testado para ações diferentes, o agente de AR treinado e testado em uma mesma ação e a estratégia *Buy and Hold* por empresa.

Empresa Treino	**Empresa Teste**	**Retorno no teste**	**Retorno na Tabela VI.6**	**Retorno *Buy and Hold*(%)**	**Significância Estatística**
Bank of America	JP Morgan	**12,45%**	11,93%	2,53%	Sim
Bank of America	Morgan Stanley	**10,49%**	9,53%	1,84%	Sim
Bank of America	Citigroup	**19,56%**	15,59%	9,93%	Não
JP Morgan	Bank of America	**10,53%**	9,46%	1,26%	Não
Morgan Stanley	Bank of America	**25,11%**	9,46%	1,26%	Não
Citigroup	Bank of America	**10,53%**	9,46%	1,26%	Sim

Observa-se que, diferentemente dos experimentos anteriores, o agente apresentou um aumento de desempenho ao ser treinado em uma empresa e avaliado em outra, os resultados da Tabela VI.8 mostram agora uma melhora consistente nos retornos obtidos pelo agente quando comparado ao cenário em que ele é simultaneamente treinado e testado na mesma ação (Tabela VI.6). Essa mudança de comportamento indica que o modelo passou a capturar padrões mais gerais de mercado, reduzindo sua dependência de características idiossincráticas de um único ativo e aumentando sua capacidade de transferência de aprendizado entre diferentes empresas.

Essa melhora pode ser atribuída a dois fatores principais. Primeiro, ao treinar em uma ação e testar em outra, o agente é exposto a uma estrutura de preços distinta daquela usada durante o aprendizado. Em vez de comprometer o desempenho — como ocorreu anteriormente — essa

divergência passou a estimular o agente a internalizar padrões de comportamento mais robustos, ligados à dinâmica setorial na totalidade e não apenas ao histórico específico de uma empresa. O resultado é um agente menos sensível a ruídos particulares de um ativo e mais responsivo a movimentos compartilhados entre instituições financeiras, como ciclos de crédito, políticas monetárias e variações macroeconômicas.

Segundo, o fato de várias empresas apresentarem níveis próximos de volatilidade, liquidez e estrutura temporal de preços permite que o conhecimento aprendido em um ativo seja aplicável a outro. Nesses casos, o agente parece ter aprendido estratégias mais genéricas de posicionamento em tendências de curto e médio prazo, beneficiando-se de maior estabilidade e menos sobreajuste quando exposto a um novo contexto.

Contudo, é importante destacar que apenas 50% dos casos apresentaram significância estatística, o que limita a confiança na interpretação de que o agente treinado em uma ação e testado em outra é, de fato, superior ao cenário em que ele é treinado e testado na mesma empresa. A ausência de significância em metade das combinações sugere que parte dos resultados positivos pode ter ocorrido por variação aleatória, o que reduz a robustez das comparações. Assim, embora alguns retornos de teste superem tanto o *buy and hold* quanto os valores obtidos na Tabela VI.6, a falta de evidência estatística consistente impede afirmar com segurança que essa abordagem generaliza melhor ou traz ganhos sistemáticos.

Em síntese, embora alguns resultados apontem para um desempenho aparentemente superior quando o agente é treinado em uma empresa e testado em outra, a ausência de significância estatística em metade dos casos impede que essa conclusão seja generalizada com segurança. Os achados sugerem um potencial de transferência de aprendizado, mas ainda insuficiente para afirmar de forma robusta que essa abordagem supera consistentemente o treinamento e teste na mesma ação. Dessa forma, os resultados devem ser interpretados com cautela, indicando a necessidade de investigações adicionais, seja ampliando o conjunto de empresas, seja reforçando os métodos de validação, para determinar se o ganho observado reflete uma capacidade real de generalização ou apenas variações conjunturais do mercado.

VI.5.3 Treinamento e teste com um portfólio de ações

O objetivo desta seção é analisar o desempenho do agente quando o treinamento e o teste são realizados em um portfólio de ações. A ideia é verificar se o modelo, ao aprender em um conjunto diversificado de ativos, consegue identificar padrões coletivos e inter-relações entre diferentes empresas, traduzindo esse conhecimento em decisões de negociação mais consistentes e robustas.

Dessa forma, o experimento consiste em treinar o agente utilizando dados históricos de um

portfólio previamente definido e, posteriormente, avaliar seu desempenho nesse mesmo portfólio. Essa metodologia é relevante porque reflete um cenário real de gestão de carteiras, no qual investidores tomam decisões considerando não apenas o comportamento isolado de uma empresa, mas a interação entre diversos ativos que compõem sua estratégia de investimento.

A comparação é feita em relação aos índices S&P500 e Nasdaq, para avaliar se o agente é capaz de superar esses benchmarks. Com isso, busca-se compreender não apenas o retorno absoluto obtido, mas também a consistência e estabilidade do agente diante da dinâmica coletiva de um portfólio de ações, elemento essencial para validar sua aplicabilidade prática em contextos reais de alocação de recursos. O portfólio é composto das seguintes ações: America Airlines, Apple, Bank of America, Cisco, Citigroup, Coca-Cola, Ford, Freport-Mcmoran, General Motors, Intel, JP Morgan, Microsoft, Morgan Stanley, Newmont, Norwegian, Pfizer, Tesla.

Tabela VI.9: Resultados comparativos entre o agente de AR sendo treinado e testado em um portfólio e índices reconhecidos do mercado.

Lucro líquido do portfólio	**Retorno do portfólio**	**Retorno S&P no período**	**Retorno Nasdaq no período**
239,79	**13,05%**	8,37%	12,07%

Os resultados obtidos com o portfólio reforçam a superioridade da estratégia do agente em relação aos principais índices de mercado. Conforme apresentado na Tabela VI.9, o agente alcançou um retorno de 13,05%, superando tanto o S&P 500 (8,37%) quanto o Nasdaq (12,07%) no mesmo período. Além disso, o lucro líquido do portfólio foi de 239,79, confirmando a capacidade da estratégia em gerar ganhos consistentes acima das métricas comparadas.

Essa diferença evidencia que o modelo não apenas acompanha o desempenho dos índices de referência, mas também os ultrapassa, demonstrando maior eficiência na captura de movimentos de valorização do mercado. Enquanto índices como S&P e Nasdaq representam médias ponderadas de um conjunto de empresas e, portanto, diluem as oscilações mais expressivas, a política de decisão do agente é ajustada continuamente às variações observadas nos preços dos ativos, o que garante maior adaptabilidade.

Dessa forma, a estratégia do agente mostra-se mais vantajosa que as abordagens tradicionais, já que alia consistência de resultados à flexibilidade de adaptação. O fato de superar tanto um índice amplo e diversificado, como o S&P, quanto um índice mais concentrado em tecnologia, como o Nasdaq, reforça a robustez da metodologia e sua aplicabilidade em diferentes contextos de mercado.

Capítulo VII Conclusões

VII.1 Análise retrospectiva

Este trabalho investigou a aplicabilidade de Algoritmos de Aprendizado por Reforço Profundo (ARP) para o problema de negociação intradiária envolvendo múltiplas ações, com foco em ambientes de alta granularidade e condições próximas ao mercado real. A principal questão norteadora consistiu em determinar se agentes baseados em ARP seriam capazes de superar estratégias tradicionais, como o *Buy and Hold*, tanto no contexto individual de cada ativo quanto na gestão conjunta de portfólios diversificados.

A abordagem proposta empregou o algoritmo Proximal Policy Optimization (PPO), selecionado após uma análise comparativa com outras arquiteturas (A2C, SAC e DDPG). Os resultados mostraram de forma consistente que o PPO ofereceu o melhor equilíbrio entre retorno e estabilidade, alcançando o maior retorno agregado (19,59%) e a menor perda acumulada (-1,37%) entre os métodos avaliados. Esses achados reforçam a adequação do PPO ao cenário dinâmico e ruidoso das operações intradiárias.

Os experimentos permitiram identificar três aspectos centrais. Primeiro, o agente demonstrou desempenho superior a benchmarks tradicionais. Quando treinado e testado na mesma ação, superou a estratégia *Buy and Hold* em todos os ativos considerados, apresentando lucros consistentes e recorrência de resultados positivos mesmo ao longo de múltiplas iterações. Em alguns casos, o ganho superou 10 pontos percentuais em relação à referência. Segundo, essa superioridade não foi apenas empírica, mas estatisticamente robusta. Testes t de Student e Mann–Whitney confirmaram diferenças significativas entre o desempenho do agente e o *Buy and Hold* para a maioria das empresas analisadas. Terceiro, a metodologia mostrou-se eficaz também ao nível de portfólio. Ao operar simultaneamente 19 ativos, o agente obteve retorno de 13,05%, sendo superior aos indicadores amplamente utilizados como S&P 500 (8,37%) e Nasdaq (12,07%) no mesmo intervalo.

Além das evidências experimentais, o estudo também contribuiu com a construção de uma infraestrutura completa para coleta, armazenamento e tratamento de dados de alta frequência provenientes do CBOE, incluindo persistência em formato Parquet e em banco de dados, permi-

tindo sua reutilização em diferentes experimentos.

Em resumo, os resultados indicam que o ARP é uma alternativa promissora para estratégias de negociação intradiária. A capacidade do agente de ajustar sua política continuamente em resposta às mudanças do ambiente representa uma vantagem competitiva importante em mercados caracterizados por alta volatilidade.

VII.2 Trabalhos futuros

Apesar dos avanços apresentados, vários caminhos para continuidade da pesquisa se mostram relevantes. Embora o agente tenha exibido desempenho consistente, sua capacidade de generalização entre ativos distintos ainda demanda investigação aprofundada, dado que a significância estatística foi alcançada apenas em cerca de metade das combinações treinadas e testadas em ações diferentes. A incorporação de variáveis externas, como indicadores macroeconômicos, notícias ou medidas de sentimento, pode ampliar a capacidade do modelo de capturar movimentos de mercado que não se refletem unicamente nos preços.

Recomenda-se também comparar o framework proposto com arquiteturas avançadas de séries temporais, como redes LSTM, Transformers ou métodos dedicados à negociação, como o Trading Deep Q-Network (TDQN), de modo a verificar se a superioridade observada do PPO persiste diante de modelos mais especializados. Além disso, a otimização de hiperparâmetros estruturais, como janelas temporais menores, estratégias de recompensa alternativas ou detecção de pontos de inflexão, podem contribuir para aprimorar o desempenho em cenários operacionais de tempo real.

Por fim, os resultados sugerem que, com aprimoramentos adicionais e expansão das arquiteturas avaliadas, agentes baseados em Aprendizado por Reforço têm potencial para consolidar-se como alternativas competitivas e eficientes para operações intradiárias em ambientes financeiros caracterizados por elevada complexidade e volatilidade.

Referências

Altuner, A. B. and Kilimci, Z. H. (2022). A novel deep reinforcement learning based stock price prediction using knowledge graph and community aware sentiments. *Turkish Journal of Electrical Engineering and Computer Sciences*, 30(4):1506 – 1524.

Atsalakis, G. and Valavanis, K. (2009). Surveying stock market forecasting techniques—part ii: Soft computing methods. *Expert Systems with Applications*, 36:5932–5941.

Carta, S., Ferreira, A., Podda, A. S., Reforgiato Recupero, D., and Sanna, A. (2021). Multi-dqn: An ensemble of deep q-learning agents for stock market forecasting. *Expert Systems with Applications*, 164.

Chou, J.-S. and Truong, T. T. H. (2019). Sliding-window metaheuristic optimization-based forecast system for foreign exchange analysis. *Soft Computing*, 23(10):3545 – 3561.

Coelho, L., Santos, A., and Costa Jr., N. (2008). Podemos prever a taxa de cambio brasileira? evidência empírica utilizando inteligência computacional e modelo econométricos. *Gestão & Produção*, 15(3):635–647.

Conegundes, L. and Pereira, A. C. M. (2020). Beating the stock market with a deep reinforcement learning day trading system.

Cui, T., Ding, S., Jin, H., and Zhang, Y. (2023). Portfolio constructions in cryptocurrency market: A cvar-based deep reinforcement learning approach. *Economic Modelling*, 119.

Elsevier (2020). Scopus: Guia de referência rápida.

Ghotbi, M. and Zahedi, M. (2024). Predicting price trends combining kinetic energy and deep reinforcement learning[formula presented]. *Expert Systems with Applications*, 244.

Gilbert, R. O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. Wiley, New York.

Gülmez, B. (2023). Stock price prediction with optimized deep lstm network with artificial rabbits optimization algorithm. *Expert Systems with Applications*, 227:120346.

Heaton, J., Polson, N., and Witte, J. (2016). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12. Available at `https://ssrn.com/abstract=2838013` or `http://dx.doi.org/10.2139/ssrn.2838013`.

Hu, Z., Zhao, Y., and Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. *Applied System Innovation*, 4(1):1 – 30.

Hull, J. C. (2014). *Options, Futures, and Other Derivatives*. Pearson, Boston, 9 edition.

Iyer, M. and Mehra, R. (2018). A survey on stock market prediction. In *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 663–668.

Jang, J. and Seong, N. (2023). Deep reinforcement learning for stock portfolio optimization by connecting with modern portfolio theory. *Expert Systems with Applications*, 218.

Kendall, M. G. (1975). *Rank Correlation Methods*. Charles Griffin, London, 4th edition.

Khan, R. H., Miah, J., Rahman, M. M., Hasan, M. M., and Mamun, M. (2023). A study of forecasting stocks price by using deep reinforcement learning. page 250 – 255. Cited by: 6.

Khushi, M. and Meng, T. L. (2019). Reinforcement learning in financial markets. *Data*, 4(3).

Kumar, G., Jain, S., and Singh, U. (2021). Stock market forecasting using computational intelligence: A survey. *Archives of Computational Methods in Engineering*, 28:1069–1101.

Kyröläinen, P. (2008). Day trading and stock price volatility. *Journal of Economics and Finance*, 32(1):75 – 89.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2016). Continuous control with deep reinforcement learning. Cited by: 3525.

Lin, C.-C., Chen, C.-S., and Chen, A.-P. (2018). Using intelligent computing and data stream mining for behavioral finance associated with market profile and financial physics. *Applied Soft Computing Journal*, 68:756 – 764.

Madhulatha, T. S. and Ghori, D. M. A. S. (2025). Deep neural network approach integrated with reinforcement learning for forecasting exchange rates using time series data and influential factors. *Scientific Reports*, 15(1). Cited by: 2; All Open Access, Gold Open Access, Green Open Access.

Maneejuk, P. and Srichaikul, W. (2021). Forecasting foreign exchange markets: further evidence using machine learning models. *Soft Computing*, 25(12):7887 – 7898.

Mann, H. B. (1945). Non-parametric tests against trend. *Econometrica*, 13:163–171.

Milana, C. and Ashta, A. (2021). Artificial intelligence techniques in finance and financial markets: A survey of the literature. *Strategic Change*, 30(3):189 – 209.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1928–1937, New York, New York, USA. PMLR.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ*, 372.

Pandolfi, M., Alastuey, A., Pérez, N., Reche, C., Castro, I., Shatalov, V., and Querol, X. (2016). Trends analysis of pm source contributions and chemical tracers in ne spain during 2004–2014: a multi-exponential approach. *Atmospheric Chemistry and Physics*, 16(18):11787–11805.

Polleit, T. (2021). Modern financial market theory – a critique based on the logic of human action. *Credit and Capital Markets*, 54(3):447 – 467.

Qiu, Y., Qiu, Y., Yuan, Y., Chen, Z., and Lee, R. (2021). Qf-tradernet: Intraday trading via deep reinforcement with quantum price levels based profit-and-loss control. *Frontiers in Artificial Intelligence*, 4.

Santos, J. O. d. and Coelho, P. A. (2010). Análise da relação risco e retorno em carteiras compostas por índices de bolsa de valores de países desenvolvidos e de países emergentes integrantes do bloco econômico bric. *Revista Contabilidade & Finanças*, 21(54):23 – 37.

Sattarov, O., Muminov, A., Lee, C. W., Kang, H. K., Oh, R., Ahn, J., Oh, H. J., and Jeon, H. S. (2020). Recommending cryptocurrency trading points with deep reinforcement learning approach. *Applied Sciences (Switzerland)*, 10(4).

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P. (2017). Trust region policy optimization.

Sun, S., Wang, R., and An, B. (2023). Reinforcement learning for quantitative trading. *ACM Transactions on Intelligent Systems and Technology*, 14(3).

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Taghavi, M., Bentahar, J., Otrok, H., and Bakhtiyari, K. (2023). A reinforcement learning model for the reliability of blockchain oracles. *Expert Systems with Applications*, 214.

Théate, T. and Ernst, D. (2021). An application of deep reinforcement learning to algorithmic trading. *Expert Systems with Applications*, 173.

Wang, C., Chen, Y., Zhang, S., and Zhang, Q. (2022a). Stock market index prediction using deep transformer model. *Expert Systems with Applications*, 208.

Wang, J., Jing, F., and He, M. (2023). Stock trading strategy of reinforcement learning driven by turning point classification. *Neural Processing Letters*, 55(3):3489 – 3508.

Wang, M.-x., Xiao, Z., Peng, H.-g., Wang, X.-k., and Wang, J.-q. (2022b). Stock price prediction for new energy vehicle enterprises: An integrated method based on time series and cloud models. *Expert Systems with Applications*, 208.

Watkins, C. J. C. H. and Dayan, P. (1992). Technical note: Q-learning. *Machine Learning*, 8:279–292.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256.

Wilmott, P. (2007). *Paul Wilmott Introduces Quantitative Finance*. John Wiley & Sons, Ltd, Chichester, 2 edition.

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

Wu, Y., Fu, Z., Liu, X., and Bing, Y. (2023). A hybrid stock market prediction model based on gng and reinforcement learning. *Expert Systems with Applications*, 228.

Zhang, W., Yin, T., Zhao, Y., Han, B., and Liu, H. (2023). Reinforcement learning for stock prediction and high-frequency trading with t+1 rules. *IEEE Access*, 11:14115 – 14127. Cited by: 5; All Open Access, Gold Open Access.

Apêndice A Sumários Estatísticos dos Conjuntos de Dados Coletados

Tabela A.1: Estatísticas descritivas das ações de empresas de tecnologia

Indicador	**Apple**	**Cisco**	**Intel**	**Microsoft**
Ações no *Top of book*				
Média	142,68	372,91	1484,66	56,15
Desvio padrão	310,50	357,99	2135,93	98,58
Mínimo	1,00	1,00	1,00	1,00
1º Quartil (25%)	30,00	200,00	613,00	5,00
Mediana (50%)	110,00	346,00	1316,00	25,00
3º Quartil (75%)	184,00	503,00	1976,00	98,00
Máximo	40.765,00	20.250,00	504.734,00	15.154,00
Preços no *Top of book*				
Média	208,07	62,68	21,49	437,46
Desvio padrão	11,35	5,21	1,62	50,36
Mínimo	149,74	5,94	12,00	342,79
1º Quartil (25%)	200,32	59,59	20,28	387,83
Mediana (50%)	208,00	63,22	21,00	447,56
3º Quartil (75%)	213,22	66,15	22,70	478,92
Máximo	244,11	189,00	70,00	600,00
Preço de uma ação				
Média	208,06	62,66	21,49	437,44
Desvio padrão	11,36	4,27	1,61	50,38
Mínimo	171,76	52,12	17,89	344,95
1º Quartil (25%)	200,32	59,58	20,28	387,80
Mediana (50%)	208,01	63,18	21,00	447,86
3º Quartil (75%)	213,22	65,97	22,68	478,86
Máximo	244,04	69,89	26,39	561,30
Número de ações vendidas				
Média	50,33	62,41	152,99	32,55
Desvio padrão	100,27	101,53	349,20	82,86
Mínimo	1,00	1,00	1,00	1,00
1º Quartil (25%)	4,00	6,00	30,00	2,00
Mediana (50%)	20,00	38,00	100,00	10,00
3º Quartil (75%)	100,00	100,00	174,00	45,00
Máximo	11.150,00	6.300,00	30.000,00	15.131,00

Tabela A.2: Estatísticas descritivas das ações de empresas do setor financeiro

Indicador	Bank of America	Citigroup	JPMorgan	Morgan Stanley
Ações no *Top of book*				
Média	1397,78	396,01	91,74	97,60
Desvio padrão	1068,93	422,73	109,97	121,01
Mínimo	1,00	1,00	1,00	1,00
1º Quartil (25%)	700,00	104,00	13,00	19,00
Mediana (50%)	1400,00	250,00	60,00	95,00
3º Quartil (75%)	2000,00	525,00	132,00	142,00
Máximo	62254,00	13019,00	10137,00	17590,00
Preços no *Top of book*				
Média	42,99	75,06	258,92	124,93
Desvio padrão	3,56	10,00	22,38	12,59
Mínimo	5,50	10,00	30,10	13,30
1º Quartil (25%)	40,97	69,21	242,39	116,41
Mediana (50%)	43,43	74,18	260,58	126,17
3º Quartil (75%)	45,37	78,61	272,28	132,50
Máximo	62,05	335,01	599,00	250,00
Preço de uma ação				
Média	43,00	75,29	258,87	125,26
Desvio padrão	3,51	9,14	21,57	11,50
Mínimo	33,08	55,54	200,97	94,54
1º Quartil (25%)	40,99	69,68	242,49	116,61
Mediana (50%)	43,40	74,21	260,66	126,31
3º Quartil (75%)	45,31	78,45	271,01	132,46
Máximo	49,30	96,90	300,95	145,45
Número de ações vendidas				
Média	123,96	54,98	35,80	42,76
Desvio padrão	258,14	95,15	57,48	56,21
Mínimo	1,00	1,00	1,00	1,00
1º Quartil (25%)	10,00	6,00	3,00	5,00
Mediana (50%)	100,00	26,00	12,00	23,00
3º Quartil (75%)	100,00	100,00	53,00	80,00
Máximo	24000,00	16497,00	4925,00	7405,00

Tabela A.3: Estatísticas descritivas das ações de empresas do setor aéreo

Indicador	**American Airlines**	**Delta Airlines**
Ações no *Top of book*		
Média	2966,02	198,13
Desvio padrão	11492,75	598,16
Mínimo	1,00	1,00
1º Quartil (25%)	1300,00	77,00
Mediana (50%)	2684,00	125,00
3º Quartil (75%)	3928,00	230,00
Máximo	1005800,00	121212,00
Preços no *Top of book*		
Média	11,17	48,06
Desvio padrão	1,11	5,34
Mínimo	6,50	27,85
1º Quartil (25%)	10,53	44,24
Mediana (50%)	11,29	48,51
3º Quartil (75%)	11,70	50,93
Máximo	50,55	129,75
Preço de uma ação		
Média	11,17	48,00
Desvio padrão	1,08	5,18
Mínimo	8,52	34,74
1º Quartil (25%)	10,54	44,28
Mediana (50%)	11,29	48,49
3º Quartil (75%)	11,69	50,81
Máximo	14,54	60,61
Número de ações vendidas		
Média	260,64	65,16
Desvio padrão	531,37	181,09
Mínimo	1,00	1,00
1º Quartil (25%)	72,00	8,00
Mediana (50%)	100,00	32,00
3º Quartil (75%)	300,00	100,00
Máximo	79088,00	24622,00

Tabela A.4: Estatísticas descritivas das ações de empresas do setor automotivo

Indicador	**Ford**	**General Motors**	**Tesla**
Ações no *Top of book*			
Média	11406,65	240,43	73,56
Desvio padrão	11545,60	282,47	217,97
Mínimo	1,00	1,00	1,00
1º Quartil (25%)	5119,00	100,00	5,00
Mediana (50%)	10093,00	200,00	25,00
3º Quartil (75%)	15500,00	301,00	100,00
Máximo	714864,00	23400,00	40101,00
Preços no *Top of book*			
Média	10,34	48,59	295,17
Desvio padrão	0,67	3,64	38,36
Mínimo	8,43	14,86	15,00
1º Quartil (25%)	9,93	46,75	263,63
Mediana (50%)	10,26	48,47	297,52
3º Quartil (75%)	10,68	50,17	326,97
Máximo	13,00	118,98	670,35
Preço de uma ação			
Média	10,34	48,55	295,16
Desvio padrão	0,67	2,75	38,34
Mínimo	8,47	41,51	214,61
1º Quartil (25%)	9,93	46,81	263,62
Mediana (50%)	10,26	48,45	297,50
3º Quartil (75%)	10,68	50,05	326,94
Máximo	11,97	54,70	366,29
Número de ações vendidas			
Média	396,41	63,72	56,47
Desvio padrão	999,02	122,57	110,88
Mínimo	1,00	1,00	1,00
1º Quartil (25%)	18,00	8,00	6,00
Mediana (50%)	100,00	43,00	25,00
3º Quartil (75%)	357,00	100,00	98,00
Máximo	62300,00	12385,00	13948,00

Tabela A.5: Estatísticas descritivas das ações de empresas do setor de mineração

Indicador	**Freeport-McMoRan**	**Newmont**
Ações no *Top of book*		
Média	246,72	333,46
Desvio padrão	301,73	588,69
Mínimo	1,00	1,00
1º Quartil (25%)	100,00	125,00
Mediana (50%)	200,00	300,00
3º Quartil (75%)	325,00	431,00
Máximo	15736,00	47890,00
Preços no *Top of book*		
Média	39,34	53,05
Desvio padrão	4,32	5,71
Mínimo	4,00	6,00
1º Quartil (25%)	37,24	48,44
Mediana (50%)	39,13	53,16
3º Quartil (75%)	41,95	57,41
Máximo	85,27	85,00
Preço de uma ação		
Média	39,30	53,12
Desvio padrão	3,90	5,31
Mínimo	27,94	41,94
1º Quartil (25%)	37,27	48,48
Mediana (50%)	39,05	53,18
3º Quartil (75%)	41,71	57,36
Máximo	48,79	66,44
Número de ações vendidas		
Média	70,59	64,66
Desvio padrão	113,05	94,52
Mínimo	1,00	1,00
1º Quartil (25%)	10,00	10,00
Mediana (50%)	52,00	43,00
3º Quartil (75%)	100,00	100,00
Máximo	10000,00	8300,00

Tabela A.6: Estatísticas descritivas das ações de empresas de setores diversos

Indicador	Coca-Cola	General Electric	Norwegian Cruise Line	Pfizer
Ações no *Top of book*				
Média	377.04	40.70	618.92	2504.75
Desvio padrão	354.65	90.28	1145.79	4711.44
Mínimo	1.00	1.00	1.00	1.00
1º Quartil (25%)	120.00	5.00	251.00	1047.00
Mediana (50%)	320.00	11.00	500.00	2108.00
3º Quartil (75%)	541.00	49.00	717.00	3400.00
Máximo	20118.00	10000.00	147871.00	861176.00
Preços no *Top of book*				
Média	70.71	222.78	19.21	24.15
Desvio padrão	1.34	30.38	2.37	1.32
Mínimo	62.30	25.00	11.00	14.57
1º Quartil (25%)	69.52	199.96	17.45	23.14
Mediana (50%)	70.91	221.91	18.91	24.14
3º Quartil (75%)	71.74	247.48	20.30	25.22
Máximo	86.88	402.25	34.33	30.36
Preço de uma ação				
Média	70.70	223.16	19.20	24.15
Desvio padrão	1.33	28.21	2.33	1.32
Mínimo	66.58	159.72	14.21	20.96
1º Quartil (25%)	69.52	200.14	17.46	23.14
Mediana (50%)	70.90	221.89	18.90	24.15
3º Quartil (75%)	71.72	247.35	20.29	25.24
Máximo	74.38	274.83	26.89	27.24
Número de ações vendidas				
Média	53.13	36.94	111.52	168.00
Desvio padrão	87.58	83.16	336.81	726.62
Mínimo	1.00	1.00	1.00	1.00
1º Quartil (25%)	3.00	4.00	15.00	10.00
Mediana (50%)	24.00	14.00	100.00	100.00
3º Quartil (75%)	100.00	59.00	100.00	170.00
Máximo	8600.00	35862.00	35531.00	374266.00