Bruno Ogata Franchi

Análise comparativa das metodologias de Markowitz, Kelly e Aprendizado por Reforço em Carteiras de Investimentos - uma abordagem computacional

Brasil

2021

Bruno Ogata Franchi

Análise comparativa das metodologias de Markowitz, Kelly e Aprendizado por Reforço em Carteiras de Investimentos - uma abordagem computacional

Trabalho de Graduação de curso apresentado ao Instituto de Ciência e Tecnologia - UNIFESP como parte das atividades para obtenção do título de Bacharel em Engenharia de Computação.

Universidade Federal de São Paulo

Instituto de Ciência e Tecnologia

Orientador: Prof. Dr. Renato César Sato

Brasil

2021

Bruno Ogata Franchi
Análise comparativa das metodologias de Markowitz, Kelly e Aprendizado por Reforço em Carteiras de Investimentos - uma abordagem computacional/ Bruno Ogata Franchi. – Brasil, 2021-
70 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Renato César Sato

TCC – Universidade Federal de São Paulo
Instituto de Ciência e Tecnologia , 2021.

1. Aprendizado por Reforço. 2. Aprendizado de Máquina. 2. Análise de Investimentos. I. Prof. Dr. Renato César Sato. II. Universidade Federal de São Paulo. III. Instituto de Ciência e Tecnologia. IV. Bacharel em Engenharia de Computação

Bruno Ogata Franchi

Análise comparativa das metodologias de Markowitz, Kelly e Aprendizado por Reforço em Carteiras de Investimentos - uma abordagem computacional

Trabalho de Graduação de curso apresentado ao Instituto de Ciência e Tecnologia - UNIFESP como parte das atividades para obtenção do título de Bacharel em Engenharia de Computação.

Trabalho aprovado. Brasil, 02 de Março de 2021:

Prof. Dr. Renato César Sato
Orientador

Prof. Dr. Sérgio Ronaldo Barro dos Santos
Professor Convidado

Profº Dra. Juliana Souza Scriptore Moreira
Professora Convidada

Prof. Dr. Tiago de Oliveira
Professor Convidado - Suplente

Brasil
2021

Dedico este trabalho a meus pais José E. Franchi e Rosa S. O. Franchi
que sempre estiveram ao meu lado dando toda estrutura emocional
para que pudesse seguir meus sonhos e objetivos

Agradecimentos

Aos meus pais, José e Rosa, que desde 2015 estiveram ao meu lado dando todo suporte e muito amor para que eu pudesse chegar até aqui. A vocês, dedico o amor incondicional eterno.

A minha namorada Jamile Novo Piveta que presenciou comigo os momentos mais turbulentos da vida acadêmica, do estágio e da vida pandêmica.

Aos meus amigos, que fizeram parte da minha jornada e me fizeram sorrir até nos momentos mais difíceis.

Aos meus colegas da Microsoft, em especial a meu mentor Samuel Masini, o qual tive oportunidade de aprender muito em meu início de carreira.

A todos os docentes da instituição, que me ensinaram não apenas o conteúdo da disciplina mas também a obter o pensamento crítico e científico frente a este amplo mundo.

Aos meus orientadores, Renato Sato, pela motivação e inspiração durante o projeto, além das valiosas informações sobre a carreira em Economia; e a professora Camila Martins, que me orientou nos primeiros anos de faculdade e me fez se interessar pela Estatística.

Resumo

Aprendizado de Máquina é o campo de estudo da inteligência artificial que estuda a descoberta automatizada de padrões e vem sendo amplamente utilizado nas pesquisas acadêmicas e nos mais diversos setores da indústria (JEWELL, 2019).A abordagem por aprendizado por reforço é uma das maneiras de se reconhecer padrões, especificamente em cenários em que não se é possível o treinamento para todas as ocasiões e o agente computacional deve aprender com sua própria experiência, (SUTTON; BARTO, 2020). No mercado financeiro a busca por maximizar a lucratividade de uma carteira de investimentos é o maior desafio dos investidores. Diferentes abordagens já foram utilizadas para análise do problema, como a divisão do risco em um sistema fechado por meio da diversificação dos ativos, como proposto por Markowitz (MARKOWITZ, 1952), ou por meio de abordagem probabilística, como muitos pesquisadores aplicam através do sistema científico de apostas proposto por John Kelly (MACLEAN; ED, 2006). O objetivo principal deste trabalho de conclusão de curso foi realizar o desenvolvimento das diferentes abordagens de otimização de carteira de investimentos propostos por Markowitz, Kelly e aprendizado por reforço. Uma análise comparativa dos retornos financeiros e computacional dos resultados obtidos, buscando verificar o desempenho por inteligência artificial frente as outras estratégias. Os algoritmos de DDPG e PPO foram utilizados para treinamento dos modelos e a implementação do projeto foi com a linguagem de programação Python (ROSSUM; DRAKE, 2009) e de suas bibliotecas disponíveis em código aberto. Séries históricas das ações presentes na Bolsa de Valores de São Paulo entre os anos de 2010 e 2019 foram coletadas através da biblioteca *yahooquery* (GUTHRIE, 2020) e utilizadas para execução dos experimentos. Os resultados obtidos mostram que a abordagem por Aprendizado por Reforço é uma estratégia eficiente para alocação de carteiras em um ambiente dinâmico como o do mercado de renda variáveis, atingindo retorno acumulado de 151,7% entre o período simulado, Junho de 2017 a Dezembro de 2019. Foi a segunda metodologia mais custosa em termos de processamento para a construção de uma única carteira, enquanto Kelly obteve os resultados mais rápidos, por volta de 14 segundos e a abordagem modificada de Markowitz obteve contabilizações de até 4000 segundos de processamento.

Palavras-chave: Aprendizado por Reforço. Aprendizado de Máquina. Análise de Investimento. Alocação de Recursos. Mercado de Ações. Teoria Moderna de Portfólio. Critério de Kelly.

Abstract

Machine Learning is a field of Artificial Intelligence that studies the automated pattern recognition from data and the whole industry and researchers are applying aiming to your business needs (JEWELL, 2019). Reinforcement learning is one of the ways that machines can recognize patterns, specifically in occasions where is not possible to learn every kind of possibility and the computational agent must to learn from your own experience (SUTTON; BARTO, 2020). In financial market the search for optimize the profitability from investments is the biggest challenge tha among all investors. Differents approachs were already used to analysis this problems, as the division of the risk between the assets through static diversification as purposes Markowitz, or through a probabilistic matter as a lot of researchers are applying the scientific bet system developed from John Kelly (MACLEAN; ED, 2006). The main goal of this work was to develop the different approachs for asset allocation purposed by Markowitz, Kelly and reinforcement learning. With the results, an comparative analysis of computational anf financial returns was made aiming to verify the performance fo artificial intelligence approch in front of the others techniques. The whole project was designed with Python (ROSSUM; DRAKE, 2009) and your open-source libraries and the dataset used was historical series from São Paulo stock market from the year of 2010 to 2019, available in public APIs. The obtained results from reinforcement learning strategy shows that AI is a efficient strategy to asset allocation in a dynamical environment as the stock market, obtaining cummulated returns of 151.7% at simulated period, 2017 June and 2019 December. It was the second most computational expensive in terms of processing for build only one investment wallet, which the modified Markowitz approach achieves until 4000 seconds of processing.

Keywords: Reinforcement Learning. Machine Learning. Investment analysis. Asset allocation. Stock markets. Portfolio Model Theory. Kelly's Criteria;

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

GMV	Mínima Variância Global
MRE	Média do Retorno Esperado
DDPG	*Deep Deterministic Policy Gradient*
PPO	*Proximal Policy Optimization*, em português, Otimização por Políticas Próximas
PDM	Processo de Decisão de Markov
PDMPO	Processo de Decisão de Markov Parcialmente Observável
RNA	Redes Neurais Artificiais

Sumário

1 Introdução

Aprendizado de máquina é o campo de estudo da inteligência artificial que estuda a descoberta automatizada de padrões em um conjunto de dados (BISHOP, 2006) e está sendo empregado nos mais variados campos científicos. De acordo com a *Accenture*, empresa multinacional de serviços profissionais, é estimado que o uso de inteligência artificial agregue ao mercado financeiro um valor de U$1.2 trilhões até 2035 (BRAND, 2018).

Uma ação é um certificado de posse de uma determinada empresa e hoje, no Brasil, a empresa brasileira Brasil Bolsa, Balcão - B3 é a responsável por fornecer a infraestrutura necessária para realização de operações de compra e venda de ativos de investimentos (B3, 2020e). Cada ação possui um preço de compra que pode variar de acordo com a expectativa dos investidores em relação a empresa, sendo alguns exemplos de fatores de influência na decisão: a perspectiva de lucro da empresa, liquidez da ação no mercado e o risco do investimento (CVM, s.d.b). O retorno financeiro de uma ação representa a valorização ou desvalorização de seu preço e o seu risco associado mensura o grau de incerteza sobre esse retorno (ELTON et al., 2014). O investidor que aloca todo o seu investimento em apenas uma ação, traz consigo todo o risco de retorno associado a tal ação. Investir em múltiplos ativos é uma estratégia que mostrou-se válida para diluição do risco do investimento (CVM, s.d.b) (ELTON et al., 2014) e, desta forma, a construção da melhor carteira de investimentos que maximiza o retorno balanceando-se ao risco associado se torna o grande desafio ao investidor.

O Modelo de Markowitz (Modelo Média-Variância) (MARKOWITZ, 1952) é uma metodologia de sistema fechado que busca sob a estratégia de diversificação da carteira os pontos ótimos e de maior eficiência, através da relação retorno esperado e risco associado, (ELTON et al., 2014). Para desenvolvimento do método, Markowitz realizou uma série de suposições sobre o comportamento do investidor e do mercado, como a racionalidade do investidor ou a eficiência sem falhas do mercado. Segundo (MANGRAM, 2013), muitos estudiosos em finanças criticam o modelo pela perspectiva de que as suposições não são adequadas ao mundo real.

Desenvolvido no contexto de apostas, o Critério de Kelly (KELLY, 1956) formula uma estratégia para alocação do tamanho do aporte em apostas, popularizado como o "Sistema Científico de Apostas". O critério utiliza das probabilidades do apostador vencer ou perder em cada partida para adequar o tamanho da fração a ser investida, impedindo que o jogador arrisque desnecessariamente parte de sua fortuna em cenários desfavoráveis. Quanto maior a probabilidade de vitória, maior será o tamanho do investimento a ser alocado.

Enquanto a abordagem de Markowitz está inserida no contexto seguro de balanceamento dos riscos, pesquisadores como Edson Thorp, (MACLEAN; ED, 2006), estudavam a aplicação do critério de Kelly para construção de carteiras de investimento, mostrando uma nova frente de leitura do mercado de renda variável, mesmo que Kelly apresente riscos maiores em suas alocações.

Algoritmos de Aprendizado por Reforço buscam, através da tentativa e erro, maximizar as recompensas de um agente inserido em um ambiente (SUTTON; BARTO, 2020), sendo a tempos referência para aplicação na área da robótica (LILLICRAP et al., 2019) e que recentemente vem ganhando seu espaço nos estudos aplicados ao mercado de ações (FISCHER, 2018), (CONEGUNDES; PEREIRA, 2020) e (H. et al., 2020).

Este trabalho buscou responder as seguintes perguntas fundamentais:

- É possível um computador aprender os padrões de flutuação das ações a partir de dados históricos, obtendo melhor rentabilidade financeira que os modelos teóricos fechados?
- É possível fazer que, através de tentativa e erro, um agente computacional construa carteiras de investimento mais rentáveis que as carteiras desenvolvidas pelos modelos de sistema fechado, ou probabilísticos como Kelly?

1.1 Motivação

A construção de carteiras de investimentos é um trabalho complexo que pode gerar grandes prejuízos quando composto inadequadamente. O modelo de Markowitz detém forte impacto teórico, entretanto não se adequa ao mundo real e é pouco aplicado em prática (MANGRAM, 2013). A abordagem de Kelly, ainda que chamada de estratégia de crescimento ótimo, pode resultar em sugestões de altos riscos pelo seu perfil de apostar nos ativos.

Visto os bons resultados obtidos na aplicação de inteligência artificial no mercado financeiro (OLIVEIRA; NOBRE; ZáRATE, 2013), (CONEGUNDES; PEREIRA, 2020) e (H. et al., 2020), o presente trabalho tem por motivação comparar o desempenho de carteiras de investimentos geradas através do Aprendizado por Reforço com os modelos de Markowitz e Critério de Kelly.

1.2 Objetivos

1.2.1 Objetivo geral

Verificar o desempenho, em termos de retorno financeiro, risco e processamento computacional, das abordagens de Markowitz, Critério de Kelly e Aprendizado por Reforço para a construção de carteiras de investimentos de múltiplos ativos.

1.2.2 Objetivos específicos

- **Aquisição dos dados:** Extração da base de dados referentes às ações presentes na Bolsa de Valores de São Paulo através da biblioteca *yahooquery*, (GUTHRIE, 2020), disponibilizada para a linguagem de programação *Python*, (ROSSUM; DRAKE, 2009);
- **Pré-Processamento:** Tratamento e modelagem do conjunto de dados obtido a fim de estruturá-los para execução dos experimentos;
- **Experimentos**:
  - Implementação dos modelos anteriormente citados utilizando a ferramenta *Python*;
  - Implementação do ambiente de treinamento do modelo por aprendizado por reforço utilizando a biblioteca *OpenAI-Gym*, (BROCKMAN et al., 2016).
  - Treinamento do modelo de aprendizado por reforço utilizando a biblioteca *Stable-Baselines*, (HILL et al., 2018).
- **Simulação e análise dos resultados**: Simulação das carteiras de investimentos geradas e análise comparativa dos resultados obtidos.

1.3 Estrutura do trabalho

O trabalho está organizado em 7 capítulos posteriores a este.

Os capítulos de 2 a 5 tem como objetivo proporcionar ao leitor a fundamentação teórica necessária para entendimento da discussão deste trabalho, iniciando com uma explicação sobre o conceito de ações e carteiras de investimento. Os Capítulos 3, 4 e 5 dedicam espaço para as noções básicas, funcionamento e aplicação das metodologias de Markowitz, Kelly e Aprendizado por reforço.

O Capítulo 6 faz referência ao processo metodológico para o desenvolvimento deste projeto, abordando os conjuntos ferramentais necessários, aquisição/pré-processamento do conjunto de dados e configuração dos experimentos. No Capítulo 9, é exposto ao leitor

os resultados obtidos do desenvolvimento, demonstrando o comportamento dos gráficos e propondo uma discussão sobre o desempenho das carteiras de investimento. O Capítulo 10 finaliza com as conclusões obtida e orienta sobre novas abordagens para trabalhos futuros.

2 Ações e Mercado Financeiro

A B3 - Brasil, Bolsa, Balcão é a empresa hoje que representa a bolsa de valores no Brasil e é responsável por fornecer a infraestrutura necessária para realização de operações em ativos, como compra e venda, para investidores. Formada em 2017 pela fusão da antiga empresa representante da bolsa de valores BMFBOVESPA S.A. com a empresa de soluções e serviços financeiros no mercado de balcão organizado Cetip (B3, 2020b), a B3 em 2015 possuía em seu sistema cerca de quinhentos mil investidores pessoas físicas, já em 2020 este número saltou para 3 milhões, (B3, 2020c).

Enquanto o mercado de investimentos em renda fixa o investidor tem a garantia do retorno em condições de baixo risco, mas caso o investidor almeje a obtenção de retornos superiores, ele pode realizar aquisições dentro do mercado de renda variável, como em ações. Quando um investidor realiza compra uma ação de uma empresa, ele passa a se tornar sócio da mesma e se beneficia da valorização da empresa e de sua ação. Em contrapartida, a empresa pode sofrer o efeito inverso e o papel de sua ação também irá se desvalorizar.

Este trabalho concentra-se no desafio da composição de uma carteira de investimentos que maximiza o retorno financeiro associado a um risco. Este capítulo tem como objetivo introduzir ao leitor os conceitos sobre ação bem como as métricas de análise utilizadas durante o projeto.

2.1 Ações

Uma ação é um certificado de posse que representa a menor parcela do capital social de empresas ou sociedades anônimas, concedendo aos acionistas todos os direitos e deveres de um sócio, no limite de suas ações adquiridas (CVM, s.d.a).

Os investidores realizam a aquisição do papel da ação, organizada como um índice composto de por siglas de quatro letras maiúsculas, indicando o nome da empresa, e um número, indicando o tipo da ação (B3, 2020a):

- Sufixo "3": Ação Ordinária - ação que proporciona o direito de voto em assembleia. Ex: PTBR3, ITUB3, AMBV3.
- Sufixo "4": Ação Preferencial - ação que proporciona prioridade para recebimento de dividendos. Ex: PTBR4, ITUB4
- Sufixo "5", "6", "7"e "8": Ações preferenciais classes A, B, C e D. Ex: VALE5

- Sufixo "34": BDRs - "*Brazilian Depositary Receipt*- Recibo Depositário Brasileiro, referindo a ações de empresas internacionais adquiridas através da B3. Ex: MSFT34.

Segundo (ELDER, 2002), a escolha de uma ação "vencedora", ou seja, a ação que terá boa rentabilidade e lucros, é mais difícil que "ouvir dicas sussurradas em uma festa barulhenta"e que o comprador deve desenvolver um conjunto de sistemas e parâmetros, além da disciplina para seguir sua estratégia, para atingir o sucesso (ELDER, 2002). A equação 2.1 representa o retorno de uma ação.

$$R_{ação} = \frac{P_{t2}}{P_{t1}} - 1 \tag{2.1}$$

Em que $R_{ação}$ é o retorno da ação, P_{t1} representa o valor do preço da ação no momento de compra e P_{t2} no momento de venda. Desta forma, caso o valor $R_{ação}$ seja maior que 1, significa que a ação gerou lucros de $R_{ação}\%$ na operação, caso contrário, terá gerado prejuízo de $R_{ação}\%$. A volatilidade do preço das ações, denominado de risco do investimento, mensura o grau de incerteza sobre o rendimento do investimento. O retorno médio de uma ação $\overline{R_{acao}}$ em um tempo t é representado pela 2.2 e o risco associado a ação, calculado através da estatística de desvio-padrão $\sigma_{R_{acao}}$, 2.3 (ELTON et al., 2014).

$$\overline{R_{acao}} = \frac{\sum^T R_{acao,t}}{T} \tag{2.2}$$

$$\sigma_{R_{acao}} = \sqrt{\frac{(\sum_i^T R_{acao,i} - \overline{R_{acao}})^2}{T}} \tag{2.3}$$

Em que $R_{acao,i}$ é o retorno da ação no dia i e T é o número de dias analisados.

$\overline{R_{acao,N}}$ e $\overline{\sigma_{acao,N}}$, representam o valor esperado do retorno e risco da ação, respectivamente, estimando o possível comportamento do preço da ação para os próximos N dias, (ELTON et al., 2014).

$$\overline{R_{acao,N}} = (1 + \overline{R_{acao}})^N - 1 \tag{2.4}$$

$$\overline{\sigma_{acao,N}} = \sigma_{acao}\sqrt{N} \tag{2.5}$$

2.1.1 Carteira de Investimento de Múltiplos Ativos

Um investidor não deve concentrar seus investimentos em apenas um único ativo, devido a existência do risco e seu impacto no retorno final caso o seu valor varie negativamente. Desta forma o investidor distribui parte dos seus investimentos entre as ações, diversificando seus retornos e diluindo seu risco, porém o desafio da composição dos

investimentos se torna mais complexo, dado que são mais variáveis para serem analisadas, como a distribuição adequada do aporte financeiro entre os ativos, (ELTON et al., 2014).

Seja R_{ai} o retorno da ação i e ω_{ai} o peso da ação i dentro da carteira, o retorno financeiro da carteira de ações R_{cart} é calculado por 2.6 e seu desvio padrão por 2.7.

$$R_{cart} = \sum_{i} R_{ai}\omega_{ai} \tag{2.6}$$

$$\sigma_{cart} = \sqrt{\omega^T Cov(R)\omega} \tag{2.7}$$

Cov é a operação de Covariância da matriz de retornos.

2.1.2 Índice da Bolsa de Valores de São Paulo - IBOVESPA

Um exemplo de carteira de ações é o Índice da Bolsa de Valores de São Paulo, de acrônimo IBOVESPA. Além de ser uma carteira teóricas, segundo a (B3, 2020d) o IBOVESPA é o principal indicador da bolsa de valores brasileira e é composta pelas ações mais negociadas dentro da B3, sendo recomposta a cada quatro meses com as ações que correspondem a cerca de 80% as operações e ao volume financeiro. A Figura 1 ilustra o gráfico do índice IBOVESPA entre os anos de 1995 a 2021. Vê-se pelo gráfico que importantes acontecimentos no Brasil impactam diretamente no comportamento do mercado de ações, como foi o início da pandemia da COVID-19 no ano de 2020 e a recessão de 2008, gerando picos decrescentes no índice.

Figura 1 – IBOVESPA entre os anos de 1995 e 2021 (B3, 2021)

3 Modelo de Markowitz

Vencedor do prêmio nobel de economia de 1990 pela sua contribuição a área de seleção de Portfólios, Harry Markowitz propõe um modelo de planejamento de carteira de investimentos em sistema fechado que se tornou uma das principais metodologias estudadas (SATCHELL; SCOWCROFT, 2003).

Para Markowitz, um investidor racional buscará sempre maximizar o retorno esperado de sua carteira, R_{cart}, e minimizar o seu risco associado, σ_{cart}, (MARKOWITZ, 1952). Segundo (ELTON et al., 2014), a diversificação da carteira de investimento em múltiplos ativos proposta por Markowitz de fato reduzia o risco do investimento, como pode ser visto na Figura 2.

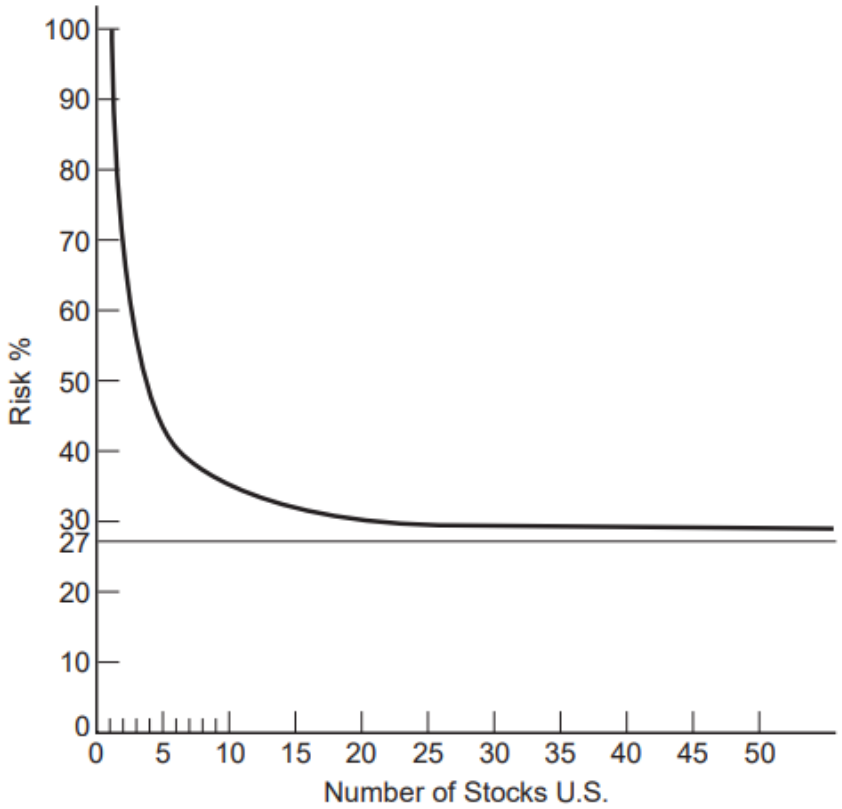


Figura 2 – Efeito do número de ativos em uma carteira ao risco do investimento (ELTON et al., 2014)

A partir dos valores de retorno esperado da carteira, R_{cart}, e do risco associado, σ_{cart}, pode-se definir a **Fronteira Eficiente** como o subconjunto das carteiras consideradas ótimos com os maiores retornos, considerando um determinado potencial de risco (ELTON et al., 2014). A Figura 3 ilustra um exemplo de uma fronteira eficiente.

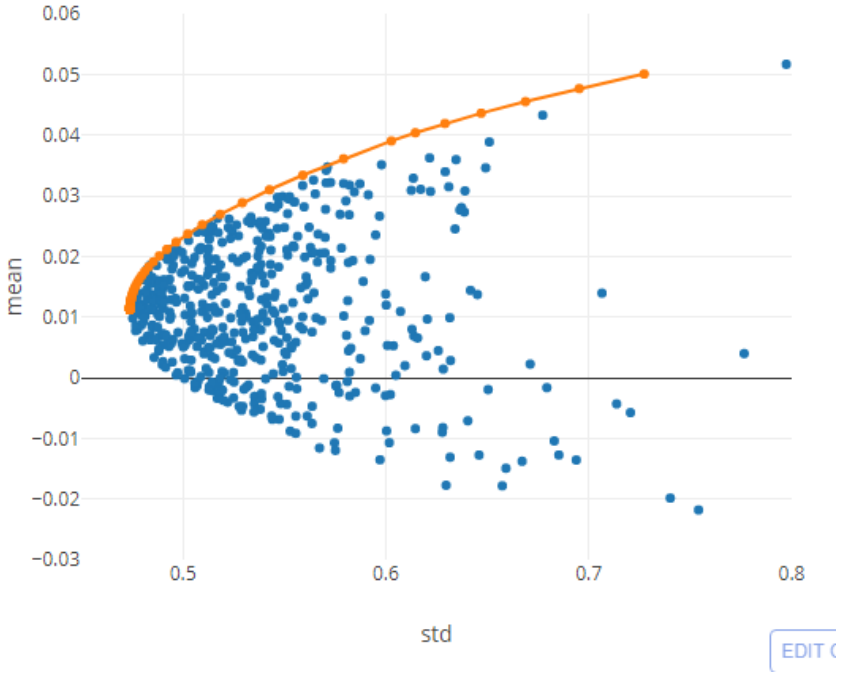


Figura 3 – Exemplo de Fronteira Eficiente (PLOTLY, 2020)

Em 1966, William F. Sharpe, que assim como Markowitz também foi laureado ao prêmio Nobel de Economia, introduziu a medida Proporção de Sharpe para analisar a performance de fundos de investimentos mútuos, a partir da relação entre seus retornos financeiros e variância (SHARPE, 1994). Essa expressão pode ser vista na equação 3.1.

$$Sharpe = \frac{R_p - r_f}{\sigma_p} \tag{3.1}$$

Em que Rp é o Retorno do Portfólio, rf a Taxa Livre de Risco e σ o risco associado da carteira.

Para Markowitz, os portfólios considerados eficientes são aqueles que:

- **GMV - Mínima Variância Global** - *Global Minimum Variance*;
- **MSR - Máxima Índice Sharpe** - *Maximum Sharpe Ratio*;

Este trabalho utilizou a abordagem pela Mínima Variância Global (GMV) para a construção da carteira de investimentos de Markowitz, selecionando a carteira de menor risco associado (σ_{cart}) como descrito pela equação 3.2.

$$Carteira_{GMV} = min([\sigma_{cart,1}, \sigma_{cart,2}, ..., \sigma_{cart,i}]) \tag{3.2}$$

Os modelos de Markowitz, pela características em sistemas fechados presumem a execução

em períodos únicos, fazendo com que o modelo funcione apenas para aquele determinado período pré-definido, assim como suas componentes.

Para o estabelecimento do modelo, Markowitz assume suposições sobre o comportamento do mercado (MARKOWITZ, 1952), (MANGRAM, 2013):

- Investidores são racionais (buscam maximizar retornos e minimizar riscos);
- Investidores estão dispostos a aceitar maiores riscos caso sejam recompensados por maiores retornos;
- Investidores recebem todas as informações pertinentes para definirem sua carteira de investimento.
- Mercados são perfeitamente eficientes;
- Inexistência de impostos ou custos de transação;

Mesmo que a contribuição dos estudos de Markowitz tenham um enorme impacto teórico, muitos estudiosos a criticam. (MANGRAM, 2013) mostra que as suposições sobre o perfil dos investidores e do mercado não se alinham ao mundo real, por exemplo: as crises que o mercado sofre por ações externas, como o cenário mais recente da pandemia, demonstram que o mercado financeiro está longe de ser eficiente. Outro exemplo são investidores que realizam aportes em ações que estão em alta no mercado, demonstrando que nem todo investidor opera sobre a racionalidade.

4 Critério de Kelly

Nascido no contexto de apostas, o Critério de Kelly, também conhecido como o "Sistema Científico de Apostas", foi desenvolvido por John Kelly como uma abordagem para dimensionamento do valor a ser aportado por um apostador em um contexto de ganha e perda. O critério utiliza das probabilidades do apostador vencer ou perder em cada partida para adequar o tamanho da fração a ser investido, impedindo que o jogador arrisque desnecessariamente parte de sua fortuna em cenários desfavoráveis. Quanto maior a probabilidade de vitória, maior será o tamanho do investimento a ser alocado. A fórmula de Kelly para apostas binárias e simples (redução total no valor aportado no caso de perda) pode ser vista na equação 4.1.

$$f_{Kelly} = p - \frac{q}{R_{vencer/perder}} \tag{4.1}$$

Em que:

- $R_{vencer/perder}$ = proporção fracionária entre lucro e prejuízo/
- p = probabilidade de se vencer;
- q = probabilidade de se perder. q = (1 - p);

Em contrapartida, no cenário do mercado de ações a situação de perda, investimento que resultou em retornos negativos ou zero, não proporciona a perda completa do investimento. Para esta ocasião, (BOCHMAN, 2018) propôs a alteração do critério de Kelly para a fórmula 4.2.

$$f_{Kelly} = \frac{p}{A} - \frac{q}{B} \tag{4.2}$$

- A = proporção fracionária de prejuízo
- B = proporção fracionária de lucros

Além disso, (BOCHMAN, 2018) mostrou que para os cenários em que se há a perda completa da ação, a 4.1 se torna uma abordagem especial da 4.2 pois: $R = \frac{B}{A}$ com A = 1, resultando em R = B.

Nesta abordagem probabilística, segundo (MACLEAN; ED, 2006), os aspectos de proporcionalidade para o dimensionamento do investimento faz com que o critério de Kelly assintoticamente maximize a taxa de retorno esperado da receita e por isso é chamada de

estratégia de crescimento ótimo. Em contrapartida, em cenários em que há domínio na probabilidade de vencer sobre perder, a fórmula pode sugerir uma grande alocação em um determinado recurso e portanto é sensível às variações a curto prazo.

Ainda segundo (MACLEAN; ED, 2006), as aplicações do critério de Kelly no mercado financeiro gerou novos desafios quando comparado a cenários de aposta convencionais, a principal delas consiste na orientação de se utilizar distribuições contínuas de probabilidade em comparação com as discretas, o que acaba gerando um grande número de valores a se analisar tornando a atividade mais complexa. Metodologias robustas para construção de Carteiras ótimas de Kelly a partir da fração de Kelly necessitam da resolução do problema de maximização quadrática (CARTA; CONVERSANO, 2020) para definição do critério ótimo dentro múltiplos critérios. (NEKRASOV, 2014) implementou um algoritmo numérico para resolução deste problema.

O critério de Kelly não atingiu grande apoio do meio acadêmico de análise de portfólio devido seu alto risco que pode ser gerado devido a orientação de grande parte dos investimentos para um mesmo recurso. Por outro lado, o critério de Kelly tem atraído personagens importantes no mundo de ações como os conceituados investidores Warren Buffett e Charlie Munger (BOCHMAN, 2018). Implementações do critério de Kelly em que há apenas o investimento parcial do resultado, alocando-se apenas uma parte do montante do que o critério define, vem sendo utilizadas como abordagens para diluir o risco e evitar a concentração da maior parte do investimento em poucos ativos.

5 Aprendizado por Reforço

Hoje a Inteligência Artificial deixou de ser lembrada apenas pelas criações artísticas como AI, Eu Robô e o familiar desenho dos anos 60 *Os Jetsons*. Para Andrew Ng, Fundador da DeepLearning.AI e Co-Fundador da gigante plataforma de ensino eletrônico Coursera, a inteligência artificial é a nova eletricidade e as técnicas de aprendizado de máquina irão impactar os mais diversos setores da indústria desde da área da saúde ao verejo (JEWELL, 2019). Aprendizado de máquina é a área de pesquisa que estuda a descoberta automatizada de padrões em um conjunto de dados, o qual se dá por abordagem supervisionada, não supervisionada ou por reforço, (BISHOP, 2006). Neste capítulo os conceitos das diferentes abordagens de aprendizado de máquina são introduzidos, adentrando-se com detalhes na principal metodologia utilizada para este trabalho, aprendizado por reforço e aprendizado por reforço profundo.

5.1 Aprendizado Supervisionado e Não-Supervisionado

Aprendizado Supervisionado e Não-Supervisionado são abordagens de aprendizado de máquina utilizados para o problema de reconhecimento de padrões dentro de um conjunto de dados em que um subconjunto deste conjunto, nomeado de Dados de Treino, é fornecido como fonte para o processo de treinamento de um modelo estatístico com objetivo da resolução do problema abordado. Segundo (BISHOP, 2006), aprendizado não-supervisionado concentra-se nos cenários em que o conjunto de dados não possuem sua variável resposta correspondente, como:

- **Análise de Agrupamento:** realiza o processo de agrupar os dados em grupos que podem ou não ser significativos com base na similaridade e relação dos valores do conjunto de dados. São exemplos: análise de segmentação para *Marketing* (HUANG; TZENG; ONG, 2007) e análise de dados sociais (MAIONE; NELSON; BARBOSA, 2019).
- **Projeção de Alta-Dimensionalidade:** aplicado a problemas de dimensão da dimensionalidade de dados, como em (MAATEN; HINTON, 2008) para visualização dos dados.

Aprendizado supervisionado busca a identificação dos padrões em um conjunto de dados que possuem sua variável resposta correspondente, sendo amplamente utilizada para resolução de problemas de classificação e regressão (BISHOP, 2006). Recentemente com o aumento na disponibilidade de dados e no aumento da força computacional, a metodologia

de aprendizado supervisionado vem atingindo bons resultados com o emprego de Redes Neurais Artificiais Profundas (AGGARWAL, 2018).

- **Classificação:** o objetivo é atribuir uma classe baseado em um conjunto finito e discreto de classes baseado nos padrões dos dados de entrada. São exemplos,
- **Regressão:** o objetivo atribuir um valor contínuo dentro de um conjunto não-finito de possibilidades, como em (OLANIYI; ADEWOLE; JIMOH, 2011) para predição de tendência de ativos.

Redes Neurais Artificiais e Aprendizagem Profunda

Segundo (AGGARWAL, 2018), Redes Neurais Artificiais são metodologias matemáticas utilizadas nos mais diversos campos da ciência e servem como modelos de simulação capazes de modelar, em teoria, qualquer função matemática dado que haja um número de dados satisfatórios. As RNAs são compostos por nós, unidades de processamento o qual são atrelados e processados os pesos de suas conexões ao sinal atingido, e estes constituem camadas dentro da rede (BISHOP, 2006). A Figura 4 ilustra o modelo de uma rede neural.

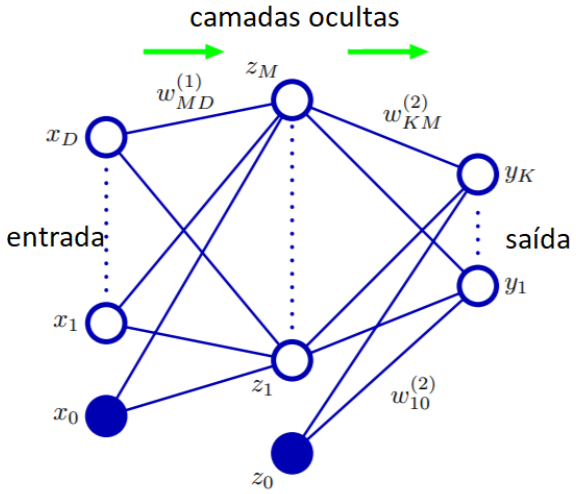
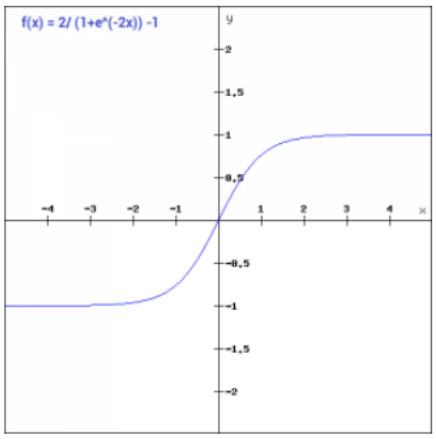


Figura 4 – Arquitetura de uma rede neural genérica - Traduzido livremente pelo autor - (BISHOP, 2006)

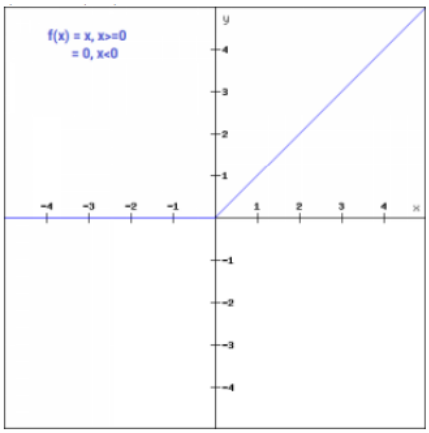
As RNAs em sua versão mais usual e básica são compostas por três camadas de nós: Entrada, Ocultas e Saída, presentes na Figura 4. A primeira é onde os valores de entradas são inicializados como o sinal da rede para que, na camada Oculta, ocorra a maior parte do processamento do sinal e ser exposta ao resultado na camada de Saída. Com o aumento no tamanho dos conjuntos de dados e o advento de tecnologias que proporcionaram máquinas de maiores poderes computacionais, o aprendizado profundo, técnica que imprime grandes redes neurais com múltiplas camadas ocultas, ganhou atenção

e começou a ser aplicada para diversos tipos de resolução de problema (AGGARWAL, 2018). Técnicas de aprendizado por reforço profundo fazem uso de RNAs profundas para cálculos de aproximação de função em ambientes de larga escala.

Em redes neurais artificiais, funções de ativação são funções matemáticas que modelam os sinais de saída da rede neural introduzindo um componente não-linear com intuito de acelerar seu aprendizado. Sem a função de ativação, a rede neural atua como um modelo de regressão linear (SHARMA; ATHAIYA, 2020). Este trabalho utilizou as funções de Tangente Hiperbólica (tanh) e a de Ativação Linear Retificada (ReLU) como função de ativação, os gráficos de suas funções podem ser vistos nas Figuras 5(a) e 5(b).



(a) Função de Ativação - tanh (SHARMA; ATHAIYA, 2020)



(b) Função de Ativação - ReLU (SHARMA; ATHAIYA, 2020)

5.2 Aprendizado por Reforço

(SUTTON; BARTO, 2020) define aprendizado por reforço como a "abordagem computacional para entendimento e automação do aprendizado orientado a objetivos". Um sistema computacional irá aprender o seu comportamento e maximizar o seu objetivo a partir de sua própria experiência em processos de tentativa e erro, modelado como um processo de Decisão de Markov.

Processos de Decisão de Markov é uma formalização de um processo decisório sequencial composto por um Agente que interage com um Ambiente através de ações (A) quando submetido a diferentes estados S do ambiente. Para cada interação de tempo t, o Ambiente sinaliza ao Agente um sinal de recompensa (R), que também pode vir como maneira de penalização, apresentando-o a uma nova situação de estado. PDMs são modelados matematicamente como $\langle S, A, P, R, \gamma \rangle$. A Figura 5 ilustra a interação entre agente e ambiente em um processo de decisão de Markov.

Nos cenários do mercado financeiro há somente a visualização parcial dos estados de processo de Markov. Nomeados de Processos de Decisão de Markov Parcialmente

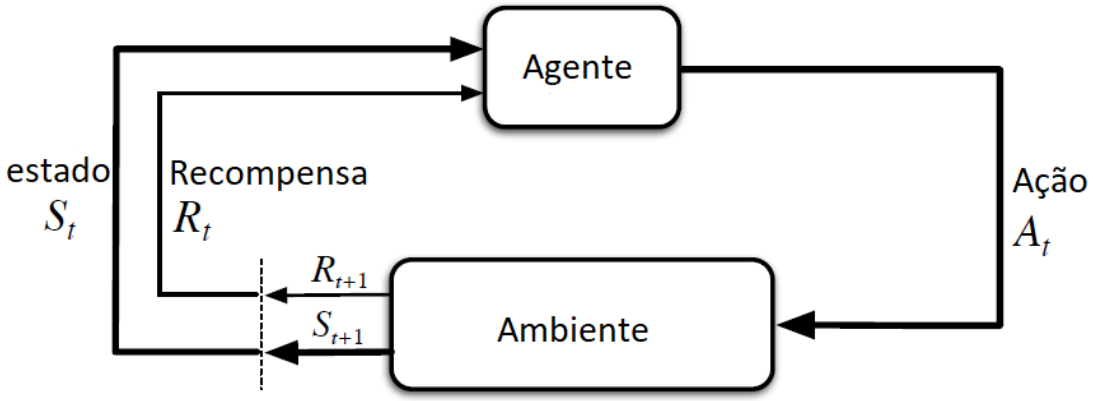


Figura 5 – Diagrama para desenvolvimento por aprendizado por Reforço - Traduzido livremente pelo autor - (SUTTON; BARTO, 2020)

Observáveis, estes cenários são uma extensão aos PDMs definidos anteriormente quando não se há visibilidade de todos os estados s do conjunto S. Para isso, o conceito de Observações é introduzido em que uma observação o representa o conjunto de probabilidades do agente estar posicionado em um estado s de S. Desta forma, o conjunto previamente definido para MDPs $\langle S, A, P, R, \gamma \rangle$ se torna $\langle S, A, O, P, R, \gamma, Z, \rangle$, em que O é o conjunto de observações o do ambiente e Z é a função de observação que mapeia os valores de observações dado os estados.

Os Processos de Decisão de Markov podem ainda ser representados em seu conjunto uma matriz de transição de probabilidade que exprime ao ambiente a probabilidade do agente atingir o S_{t+1} a partir do estado S_t, exprimindo a dinâmica de funcionamento do ambiente.

Portanto, o problema geral do Processo de Decisão de Markov consiste em entender o modelo de funcionamento do ambiente, de modo a maximizar o acúmulo de recompensas a longo prazo, enquanto apoia-se na otimização das funções de Estado-Valor e Ação-Valor de um ambiente, representados das maneiras mais simplificadas pelas equações 5.1 e 5.2.

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s] \tag{5.1}$$

(SILVER, 2015a)

$$q_\pi(s) = E_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a] \tag{5.2}$$

(SILVER, 2015a)

π, presente nas equações acima, é o conjunto de políticas o qual o Agente utiliza para atingir seus objetivos. A política mapeia a distribuição das ações a serem executadas pelo Agente em cada estado S_t.

γ representa o coeficiente de desconto dos retornos (SILVER, 2015a), este coeficiente tem o objetivo de imprimir a importância dos retornos de estados que estão a longo prazo, fazendo com que o agente pondere a próxima ação baseado na expectativa do próximo retorno.

Segundo (SUTTON; BARTO, 2020), o aprendizado por reforço faz uso da parte ferramental do PDM, a fim de formalizar a interação entre agente e ambiente. A Maximização das funções Estado-Valor e Ação-Valor é o objetivo principal, sendo o foco maior da abordagem na resolução dos cenários em que não se tem a dinâmica do ambiente explícita, ou seja, não há uma matriz de probabilidades (OTTERLO; WIERING, 2012). Nesta linha, as abordagens para resolução dos problemas dos PDMs se dividem em duas linhas: Resoluções baseadas em modelos e livre de modelos.

5.2.1 O desafio da Exploração x Explotação

Segundo (SUTTON; BARTO, 2020), um dos maiores desafios do aprendizado por reforço é o balanço do comportamento do agente em exploração ou explotação. A operação de explotar (do inglês, *exploit*) significa o exagero na escolha das ações que resultam em boas recompensas e já foram conhecidas no passado. Porém, a realização apenas de ações já conhecidas pode impedir que o agente não conheça situações que o proporcionam recompensas superiores ou máximas. Neste caso, a operação de explorar é importante de maneira que faça o agente se arriscar para atingir estados de maior recompensa no futuro. A escolha do balanço da definição de quando e quanto o agente deve explorar (ou explotar) é um dos desafios de aprendizado por reforço que, segundo (SUTTON; BARTO, 2020), ainda permanece não resolvido.

5.2.2 *Resoluções baseada em modelo e livre de modelos*

Os cenários de Aprendizado por Reforço podem envolver uma Matriz de Transição de Probabilidade que manifesta ao ambiente quais são as probabilidades de transição entre os estados. Nesses casos os cenários são chamados de resolução baseada em modelo, justamente por terem referência probabilística que rege a dinâmica do sistema e poderem ser resolvidos nestes casos os cenários são chamados de resolução baseada em modelo e podem ser resolvidos por algoritmos de Programação Dinâmica. Os ambientes que não possuem a Matriz de Transição explícita são chamados de resolução livre de modelo, (OTTERLO; WIERING, 2012).

Durante os anos de estudo no meio acadêmico, diversas metodologias foram desenvolvidas para otimizarem os resultados e explorarem as aberturas que surgiam nos avanços de Aprendizado por Reforço resolução livre de modelo, as duas principais abordagens que ganharam destaque são: Otimização das Políticas e *Q-Learning*. A Figura 6 ilustra a

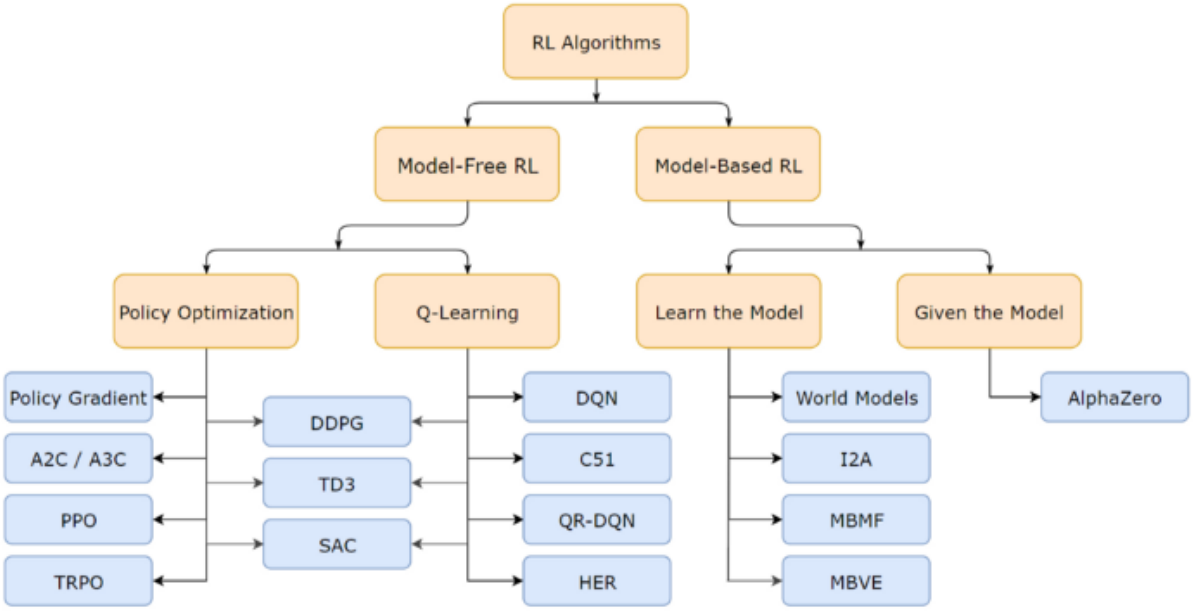


Figura 6 – Taxonomia de uma amostra de algoritmos de Aprendizado por Reforço (BROCKMAN et al., 2016)

Taxonomia de uma amostra de algoritmos de Aprendizado por Reforço.

Cada algoritmo possui a sua particularidade e maneira de resolução do problema: algoritmos como SARSA avaliam e melhoram a política utilizada (SILVER, 2015b) enquanto algoritmos como SAC atuam e aprimoram uma política diferente da selecionadora de ações (HAARNOJA et al., 2018). Algoritmos como SARSA são chamados de métodos *On-Policy* enquanto os métodos *Off-Policy* analisam uma política externa para aprendizado.

5.2.3 Métodos por Aproximação de Função

Em cenários de grande porte em que o espaço de estados possui um grande número de estados disponíveis ou que o mesma seja contínuo, a operação de calcular a função Estado-Valor para cada estado se torna uma abordagem inviável em lidar com o problema e não-escalável.

(SUTTON; BARTO, 2020), a abordagem por Aproximação de Função resolve este problema introduzindo o conceito de utilizar um aproximador para estimar os pesos da função Valor que melhor representa o estado original. Com aproximadores de função, não é necessário que o agente execute o trabalho árduo de calcular todos os valores da função Valor, realizando a operação diretamente com os pesos estimados para obtenção. Frequentemente, o uso de aprendizado supervisionado é utilizado como aproximadores de função, o que deu origem a abordagem aprendizado por reforço profundo, em que se faz uso de redes neurais artificiais profundas como aproximadores de função.

Segundo (SUTTON; BARTO, 2020), os métodos de Aprendizado por Reforço por aproximação de função também são aplicáveis a modelos de Processos de Decisão de Markov Parcialmente Observáveis.

5.2.4 *Deep Deterministic Policy Gradient*

O *Deep Deterministic Policy Gradient* - DDPG é um algoritmo *Off-Policy* por resolução livre de modelo de Aprendizado por Reforço Profundo desenvolvido por (LILLICRAP et al., 2019) para aplicação de Aprendizado por Reforço em espaços de observações e ações contínuos e de alta dimensionalidade. É um método por análise de gradiente que deterministicamente otimiza suas funções objetivas. Sua inspiração foi a de assimilar as ideias principais propostos pelos algoritmos de DPG e DQN para operarem em espaços de ações contínuos

O método faz uso de duas redes neurais, ator e crítica, para elaboração dos resultados (comuns a técnicas nomeadas de Ator-Crítica, como A2C e PPO). A primeira é treinada para ser responsável por propor uma ação em um determinado estado e a segunda para predizer a qualidade de uma ação em um determinado estado e ação. Para apoiar ao treinamento das redes, DDPG utiliza de um *Buffer* de Experiência para armazenar amostras durante a execução do treinamento (ideia utilizada no algoritmo DQN - *Deep Q Learning*). A Figura 7 demonstra o pseudocódigo da metodologia de DDPG (LILLICRAP et al., 2019).

5.2.5 Otimização por Políticas Próximas - PPO

O método por Otimização por Políticas Próximas (do inglês, *Proximal Policy Optimization*) é um conjunto de algoritmos por análise de gradiente do espaço de políticas, *On-Policy* por resolução livre de modelo e Ator-Crítica que busca alternar entre amostragem dos dados através da interação com o ambiente e otimizar as funções objetivas estocasticamente para o a otimização da política d explorar e explotar (SCHULMAN et al., 2017).

A ideia principal do algoritmo é que o resultado da próxima política não esteja longe da anterior. Conforme o agente vai adquirindo experiência e a política estocástica adquire mais volume de conhecimento, a distribuição das probabilidades das ações se tornam menos aleatórias e ação explotadas começam a ser prioridade para a política. Esta tendência a longo prazo de explotar pode fazer com que a política presa a um ponto ótimo local, não se permitindo a obter resultados maiores através da ação de explorar (**??**).

Há duas implementações principais da PPO: PPO por penalidade e PPO por recorte. Em que a primeira utiliza de ferramentais estatísticos para cálculo de divergência de distribuições probabilísticas (Divergência de Kullback-Leibler) para atualização da

Algorithm 1 DDPG algorithm

Randomly initialize critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ.
Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer R
for episode = 1, M **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial observation state s_1
  for t = 1, T **do**
    Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise
    Execute action a_t and observe reward r_t and observe new state s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in R
    Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from R
    Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$
    Update critic by minimizing the loss: $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$
    Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i}$$

    Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$

  end for
end for

Figura 7 – Pseudocódigo - DDPG (LILLICRAP et al., 2019)

política. A PPO por recorte não utiliza tal ferramental, realizando sua otimização das políticas apenas pela otimização do Gradiente. A Figura 8 demonstra o pseudocódigo da PPO por recorte (OPENAI,).

5.2.6 Aprendizado por Reforço em comparação com as outras formas de aprendizado

Para (SUTTON; BARTO, 2020), aprendizado por reforço é diferente das outras abordagens de aprendizado de máquina. Enquanto o aprendizado supervisionado busca generalizar a atribuição das classes corretas para novos dados de entrada, baseado nos padrões pré-estabelecidos durante a fase de treinamento, o modelo por reforço adquire aprendizado através do processo de tentativa e erro em ambiente interativos. Neste tipo de cenário, não é frequente a obtenção de um conjunto de dados suficiente que exprima o comportamento adequado que o modelo deve tomar para atingir o sucesso, sendo necessário que o sistema computacional aprenda com a sua experiência no ambiente.

Em comparação ao aprendizado não-supervisionado, (SUTTON; BARTO, 2020)

Algorithm 1 PPO-Clip

Input: initial policy parameters θ_0, initial value function parameters ϕ_0
for $k = 0, 1, 2, ...$ **do**
  Collect set of trajectories $\mathcal{D}_k = \{\tau_i\}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment.
  Compute rewards-to-go $\hat{R}_t$.
  Compute advantage estimates, $\hat{A}_t$ (using any method of advantage estimation) based on the current value function V_{ϕ_k}.
  Update the policy by maximizing the PPO-Clip objective:

$$\theta_{k+1} = \arg\max_{\theta} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^{\pi_{\theta_k}}(s_t, a_t), \quad g(\epsilon, A^{\pi_{\theta_k}}(s_t, a_t)) \right),$$

  typically via stochastic gradient ascent with Adam.
  Fit value function by regression on mean-squared error:

$$\phi_{k+1} = \arg\min_{\phi} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \left(V_\phi(s_t) - \hat{R}_t \right)^2,$$

  typically via some gradient descent algorithm.
end for

Figura 8 – Pseudocódigo - PPO (OPENAI,)

ressalta que, mesmo que ambas abordagens não se utilizem de um conjunto de dados com a variável resposta explicitamente fornecida, ambas metodologias são diferentes em quesito do objetivo a ser atingido: na primeira, o objetivo concentra-se na identificação de padrões e estruturas ocultas nos dados, enquanto o aprendizado por reforço busca maximizar a obtenção de recompensas por um agente computacional em um meio interativo.

Além disso, o desafio de Exploração e Explotação não surge aos outros paradigmas, sendo necessário o tratamento dos casos e a análise da viabilidade do balanço restrito a apenas problemas de aprendizado por reforço.

5.2.7 Aprendizado de Máquina aplicado a finanças

No campo de finanças, o uso de técnicas de aprendizado de máquina são cada vez mais empregadas nas pesquisas principalmente sob utilização do paradigma supervisionado (FISCHER, 2018). (OLIVEIRA; NOBRE; ZáRATE, 2013) publicaram um artigo nomeado "*Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index - Case study of PETR4 Petrobras, Brazil*", considerando as flutuações do ativo na bolsa de valores cujo objetivo era analisar as flutuações do ativo na bolsa de valores e verificar quais eram as principais variáveis que impactavam diretamente no preço da ação. O estudo concluiu que o uso de inteligência artificial, em específico Redes Neurais Artificiais, com finalidade de prever o movimento de ações é uma alternativa

válida para os métodos tradicionais.

Ainda em sua pesquisa (FISCHER, 2018) mostra que mesmo que aprendizado supervisionado seja o principal paradigma no mercado financeiro, o uso de aprendizado por reforço vem adquirindo seu espaço e está se tornando uma área de estudo cada vez mais popular e com bons resultados. (DENG et al., 2017) propuseram o uso de aprendizado por reforço em seu trabalho *Deep Direct Reinforcement Learning for Financial Signal Representation* para construção de um sistema inteligente, aplicado ao processamento de senais financeiros em tempo-real, propondo um modelo inspirado nos conceitos de aprendizado de reforço profundo. Em outro projeto, (EILERS et al., 2014) aplicaram em uma abordagem utilizando Aprendizado por Reforço para o desenvolvimento de algoritmos para apoio a decisão no mercado financeiro em cenários de estratégia de sazonalidade.

6 Métodos

Neste capítulo será apresentada a metodologia utilizada no desenvolvimento do projeto.

O desenvolvimento deste trabalho envolveu o processo de experimentação com base de dados históricos de ações da Bolsa de Valores de São Paulo e aplicação das metodologias de Markowitz, Kelly e *Aprendizado por Reforço*, computando e analisando as métricas financeiras e computacionais para seleção dos portfólios. As sete etapas do desenvolvimento do projeto podem ser vistas junto da descrição das tarefas a seguir:

- **Etapa 1** - Aquisição dos Dados;
- **Etapa 2** - Pré-Processamento dos Dados;
- **Etapa 3** - Análise Exploratória dos Dados;
- **Etapa 4** - Experimentação com Regras de Markowitz;
- **Etapa 5** - Experimentação com Critério de Kelly;
- **Etapa 6** - Treinamento e Experimentação do Modelo de Aprendizado por Reforço
- **Etapa 7** - Análise dos Resultados

Primeiramente, descreve-se com detalhes o conjunto ferramental, bem como as versões utilizadas e o aspecto físico *hardware* envolvido.

Na segunda seção, expõe-se Fluxo Lógico das Atividades executadas é apresentado, descrevendo as tarefas aplicadas em cada etapa como a Aquisição, Pré-Processamento e análise exploratória do conjunto de dados, além da etapa de desenvolvimento dos *scripts* de experimentação e análise performática.

Por último, são definidas as métricas de performance financeiras e computacionais para avaliação dos resultados de cada carteira e apresentado dos portfólios de *benchmark* utilizados para comparação com os resultados obtidos.

6.1 Conjunto Ferramental

6.1.1 *Softwares*

Todo o trabalho computacional, desde a modelagem dos dados até a construção de visuais gráficos, foi feito utilizando a linguagem de programação *Python* devido sua

versatilidade quanto plataforma, além de possuir uma comunidade unida e colaborativa em desenvolvimento de projetos *open-source* e de pesquisas para cunho científico (ROSSUM; DRAKE, 2009), além da ferramenta de *notebooks* interativos Jupyter.

Para cada etapa do desenvolvimento do trabalho, diversas bibliotecas eram adicionadas para fins de resolução de problemas conhecidos e já implementados pela comunidade, a Tabela 6.1.1 mostra o relativo do pacote e sua versão utilizada. Com o intuito de extrair os dados das ações da Bolsa de Valores de São Paulo - BOVESPA durante a Etapa 1 - Aquisição de Dados, implementou-se um código em Python utilizando a biblioteca yahooquery. Esta biblioteca é desenvolvida e mantida pela comunidade e funciona como uma interface de comunicação com a fonte de dados financeiro do Yahoo Finance através de API.

Para a Etapa 4 - Treinamento do Modelo de Aprendizado por Reforço, as ferramentas *OpenAI Gym* e *Stable-Baselines* foram utilizadas para a construção do ambiente de treinamento e definição do modelo de Aprendizado de Máquina, respectivamente; o*MatplotLib* foi utilizada durante todas etapas do desenvolvimento por se tratar de um conhecido pacote para criação de gráficos e visuais, porém, o seu maior uso foi emplacado durante as Etapa 2 - Análise Exploratória dos Dados e Etapa 5 - Análise dos Resultados. Em todas as etapas os pacotes **Pandas**, para fornecimento de estruturas de dados para fins estatísticos, e *NumPy*, para de poderosas funções matemáticas, foram utilizados massivamente.

Software	**Versão**
Python (ROSSUM; DRAKE, 2009)	3.6.8
Pandas (MCKINNEY, 2010)	1.1.5
NumPy (HARRIS et al., 2020)	1.19.5
Stable-Baselines (HILL et al., 2018)	2.10.1
OpenAI Gym (BROCKMAN et al., 2016)	0.18.0
MatplotLib (HUNTER, 2007)	3.3.4
Yahooquery (GUTHRIE, 2020)	2.2.8

Tabela 1 – Tabela com a versão dos *softwares*

6.1.2 Configurações da Máquina

A configuração da máquina utilizada para execução dos experimentos pode ser visto na Tabela 6.1.2

Processador:	Intel(R) Core(TM) i5-9300H CPU @ 2.40GHz 2.40GHz
Placa Gráfica:	NVIDIA GeForce GTX 1650
Memória RAM:	16 GB
SSD	256 GB
Sistema Operacional:	Windows 10 Home - 64 bits

Tabela 2 – Tabela com as configurações da máquina utilizada para desenvolvimento

Uma observação a ser feita quanto ao sistema operacional é que, para a execução dos *scripts* de treinamento do modelo de Aprendizado de Máquina, foi utilizado o *software Windows Subsystem for Linux* executando a distribuição Ubuntu 24.04.1 LTS.

6.2 *Benchmark*

Dados históricos do IBOVESPA-Índice da Bolsa de Valores de São Paulo e do IPCA-Índice Nacional de Preços ao Consumidor Amplo são utilizados a fim de realizar a análise comparativa dos desempenhos financeiros das carteiras de investimentos. O IBOVESPA é o principal indicador de desempenho das ações negociadas na B3 e é resultado de uma carteira de ações que reúne os principais ativos da bolsa de valores brasileira (B3). Neste trabalho o IBOVESPA representará a taxa de valorização do mercado e o IPCA representará a medida de inflação dos preços ao consumidor. As Figuras 9 ilustra a flutuação dos preços do IBOVESPA e IPCA e 10 demonstra a valorização financeira da carteira teórica do IBOVESPA entre os anos de 2012 a 2019.

As tabelas de IPCA e IBOVESPA foram construídas através da coleta operacional por parte do autor diretamente das fontes de dados disponíveis respectivamente em <http://indiceseconomicos.secovi.com.br/indicadormensal.php?idindicador=61#> e <https://www.investing.com/indices/bovespa-historical-data>.

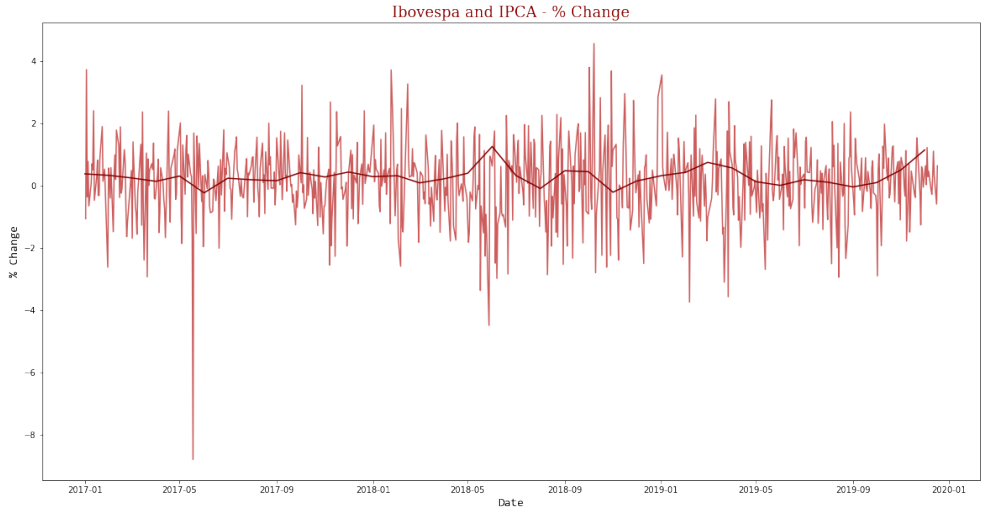


Figura 9 – Variação Percentual dos Índices do IBOVESPA e IPCA entre 2017 e 2019

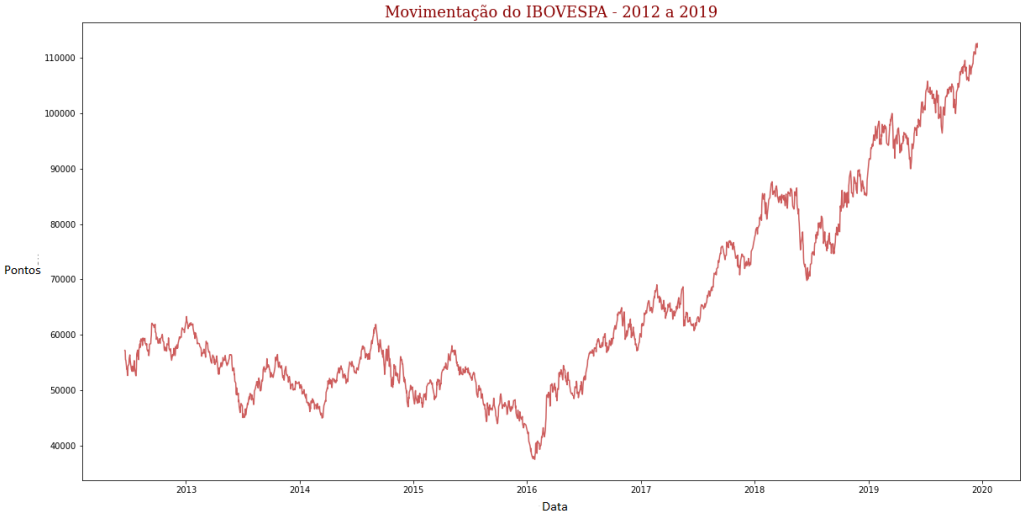


Figura 10 – Gráfico da valorização financeira do IBOVESPA entre os anos de 2012 a 2019

6.3 Conjunto de Dados

Neste capítulo são descritos as rotinas de aquisição e pré-processamento dos dados, preparando-os para as etapas de experimentação.

6.3.1 Aquisição dos Dados

Para aquisição dos dados foi necessário o desenvolvimento de *script* Python que iterava sobre os índices presentes na B3, realizando chamadas via API através do pacote Yahooquery (descrito no capítulo anterior). Através do *script*, fora construída uma base de dados de 10 anos históricos referente a 244 ativos e o conjunto de dados bruto final ficou composto por 439269 linhas e 10 colunas. A Tabela 3 demonstra uma amostra de duas linhas do conjunto final, a descrição de cada coluna segue abaixo.

Volume	**Close**	**Dividends**	**High**	**Adj Close**	**Splits**	**Low**
273275.0	9.092740	0.0	9.144280	5.694830	NaN	8.835050
4516900.0	47.50000	0.0	49.02999	46.621166	NaN	46.950001
Symbol	**Open**	**Date**				
ABCB4.SA	8.835050	2010-01-04				
YDUQ3.SA	48.509998	2019-12-30				

Tabela 3 – Tabela Amostra do Conjunto de Dados após a Etapa 1

- ***Volume:*** Volume financeiro do ativo negociado durante o pregão;
- ***Close:*** Preço do ativo após o fechamento do pregão;
- ***Dividends:*** Valor distribuido em dividendos aos acionistas;
- ***High:*** Maior preço atingido pelo ativo durante o pregão;
- ***Adj Close:*** Preço de fechamento ajustado após o fim do pregão;
- ***Splits:*** Divisão proporcional do ativo após uma ação de divisão pela empresa;
- ***Low:*** Menor preço atingido pelo ativo durante o pregão;
- ***Open:*** Preço do ativo no início do pregão;
- ***Symbol:*** Índice do ativo;
- ***Date:*** Data do pregão

6.4 Pré-Processamento dos Dados

Antes do início da construção dos experimentos e execução, foi executado um pré-processamento nos dados brutos obtidos. Para (AALST, 2016), o sucesso para o processo de ciência de dados está totalmente atrelado a qualidade dos dados. Caso não houver confiança neles, os resultados como todo o processo se tornam menos valiosos.

Para este trabalho, o processo de limpeza dos dados foi a etapa o qual mais exigiu tempo e esforço. Após o processo de limpeza, o conjunto final para execução dos

experimentos ficou composto por 295 mil linhas e 8 colunas, aproximadamente 68% da tabela bruta, que serviu como base para a modelagem de cada modelo de dados específicos para as metodologias utilizadas para análise.

Análise de Qualidade e Limpeza dos Dados

Após a obtenção dos dados brutos, fora aplicada uma análise exploratória dos dados e desenvolvido códigos em Python para adequar a propriedade dos dados. Alguns dados vieram com valores nulos ou incompletos, como pode ser visto na Tabela 3 nas colunas de *Split* e *Dividends*, e tiveram de ser tratados para execução das análises. Outro aspecto dos dados coletados é o desbalanço na quantidade de dados obtidos por índice e por data, como podem ser vistos nas Figuras 9 e 10. A Figura 12 ilustra a distribuição dos dados obtidos por índice de ação e demonstra que parte dos ativos não possuem informações referentes a todo o intervalo de tempo de 2010 a 2019. Esta interpretação motivou uma filtragem nos índices que possuíam menos de 75% dos dados, resultando em 152 ativos. Na Figura 11, o qual se obtém a distribuição dos dados obtidos por data, vê-se que há um aspecto crescente na distribuição. Foi observado que a partir do dia 15 de Junho de 2012 os dados se mantinham consistentes, motivando a filtragem dos dados referentes ao intervalo de tempo anterior.

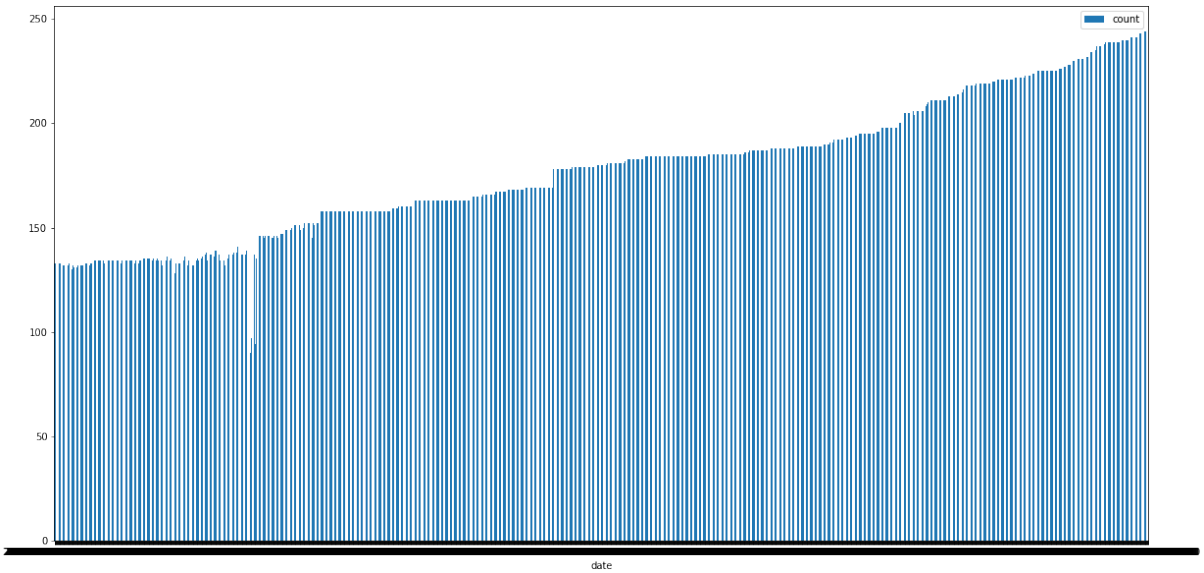


Figura 11 – Distribuição dos Dados com relação à variável *Date*

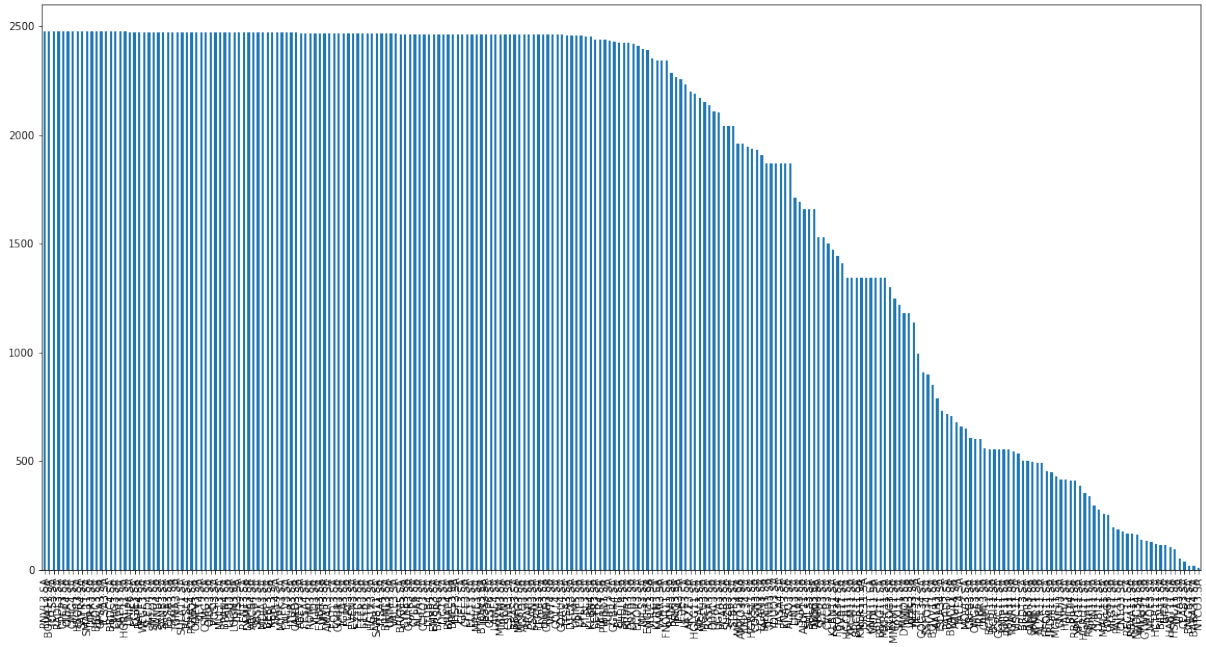


Figura 12 – Distribuição dos Dados com relação à variável *Symbol*

Após a finalização das análises, o conjunto de dados resultante após o primeiro *check* de qualidade ficou composto por 295144 linhas e 8 colunas (exclusão das colunas *Dividends* e *Splits*), como pode ser visto na tabela 6.4. Para este conjunto de dados, o intervalo histórico é de **19 de Junho de 2012 a 30 de Dezembro de 2019**, *clean_data*.

Conjunto de Dados	**Número de Linhas**	**Número de Colunas**
Bruto	439269	10
Pós 1º Filtragem	*(-32%)* 295144	8

Tabela 4 – Dimensões dos Conjuntos de Dados Bruto e após aplicação da 1º Filtragem

6.4.1 Cálculo de Retorno Financeiro e Análise de Correlação

Com os conjuntos de dados divididos, calculou-se o retorno financeiro (equação 2.1) diário de cada ação em relação ao dia anterior, incrementado-se uma nova coluna. Note que não seria possível calcular-se o retorno diário referente ao primeiro dia da tabela por não se ter o dia anterior a este, portanto, as linhas referentes ao primeiro pregão foram removidas. A partir nova coluna Retorno e utilizando o conjunto de dados de Treino, foi construída uma segunda tabela, 5, constituída apenas dos retornos diários de cada ação.

ABCB4.SA	TIET4.SA	ALSO3.SA	...	VULC3.SA	WEGE3.SA	YDUQ3.SA
0.0225	0.0025	0.0308		0.000	-0.0162	0.0166
-0.0450	0.000	-0.0316		0.000	0.0074	-0.0184
-0.0317	0.0057	-0.0236		0.0452	0.0036	0.0146

Tabela 5 – Amostra da tabela de retornos financeiros

Ações com alta correlação entre seus retornos possuem flutuações de preço semelhantes. Alocar ações deste perfil na carteira de investimento pode aumentar o seu risco associado, já que comportamento similar do retorno é similar - se o preço de uma ação cair as chances da outra seguir o mesmo movimento são altas. Com o interesse de filtrar os ativos que possuem alta correlação entre si foi computado a partir da Tabela de Retorno 5 a matriz de correlação entre os ativos e selecionados os pares de ativos que possuíam correlação maior que 0.85, para cada par o ativo com menor retorno financeiro era selecionado para ser removido do conjunto. A Tabela 6.4.1 demonstra os pares de alta correlação.

Pares de Correlação >0.85				
Índice	**M. E. R. (252)**	**Corr**	**Índice**	**M. E. R. (252)**
BBDC4.SA	0.1944	**0.8975**	BBDC3.SA	0.2250
GGBR4.SA	0.0108	**0.85384**	GGBR3.SA	0.0557
GOAU4.SA	-01365	**0.87696**	GGBR4.SA	0.0108
ITUB4.SA	0.2313	**0.91677**	ITSA4.SA	0.1958
ITUB4.SA	0.2313	**0.87212**	ITUB3.SA	0.2355
PETR4.SA	0.0643	**0.94997**	PETR3.SA	0.0627
VALE3.SA	0.1016	**0.89973**	BRAP3.SA	0.0713

Tabela 6 – Tabela composta dos pares de ativos com correlação maior que 0.85

Portanto, sete ativos (BBDC4.SA, BRAP4.SA, GGBR4.SA, GOAU4.SA, ITSA4.SA, ITUB4.SA, PETR3.SA) foram filtrados do conjuntos de dados finalizando-o com o total de 151 índices. Para uma visão geral do conjunto de dados resultante, análises descritivas referentes a média, desvio padrão, máximo e mínimo foram calculados para cada ano, os resultados podem ser vistos abaixo nas Tabelas 7, 8, 9, 10, 11, 12, 13 e 14.

*M. E. R. (252) = Média Esperada do Retorno para 252 dias. *D.P. = Desvio Padrão.

2012	**Volume**	**Close**	**High**	**Adj Close**	**Low**	**Open**	**Returns**
Média	1.54 x 10^6	40.34	40.98	36.75	39.65	40.38	0.0012
D. P.	5.24 x 10^6	122.43	125.33	122.84	119.55	122.81	0.025
Mínimo	0.00	0.25	0.195	-0.007	0.25	0.27	-0.329
Máximo	3.05 x 10^8	1398.00	1398.00	1398.00	1352.00	1391.0	0.818

Tabela 7 – Estatísticas descritivas do conjunto de dados filtrado referente ao período de Junho a Dezembro de 2012.

2013	**Volume**	**Close**	**High**	**Adj Close**	**Low**	**Open**	**Returns**
Média	1.92 x 10^6	30.83	31.35	27.49	30.37	30.90	0.0001
D. P.	5.11 x 10^7	65.84	67.37	65.95	64.68	66.26	0.034
Mínimo	0.00	0.13	0.14	-0.0073	0.12	0.12	-0.40
Máximo	1.67 x 10^9	1025.00	1094.00	1025.00	1019.00	1084.00	4.125

Tabela 8 – Estatísticas descritivas do conjunto de dados filtrado referente ao ano de 2013.

2014	**Volume**	**Close**	**High**	**Adj Close**	**Low**	**Open**	**Returns**
Média	2.29 x 10^6	23.70	24.00	21.00	23.41	23.72	-0.0002
D. P.	1.87 x 10^7	41.59	41.73	41.41	41.41	41.58	0.029
Mínimo	0.00	0.14	0.14	-0.014	0.14	0.14	-0.491
Máximo	2.90 x 10^9	803.77	803.77	803.77	803.77	803.77	1.00

Tabela 9 – Estatísticas descritivas do conjunto de dados filtrado referente ao ano de 2014.

2015	**Volume**	**Close**	**High**	**Adj Close**	**Low**	**Open**	**Returns**
Média	2.22 x 10^6	23.51	23.76	21.48	23.25	23.51	-0.0004
D. P.	8.93 x 10^6	74.16	74.24	73.99	74.04	74.15	0.035
Mínimo	0.00	0.03	0.031	0.038	0.03	0.031	-0.54
Máximo	5.59 x 10^8	1327.28	1327.28	1327.28	1327.28	1327.28	1.88

Tabela 10 – Estatísticas descritivas do conjunto de dados filtrado referente ao ano de 2015.

2016	**Volume**	**Close**	**High**	**Adj Close**	**Low**	**Open**	**Returns**
Média	3.04 x 10^6	27.03	27.54	25.64	27.01	27.27	-0.001
D. P.	2.42 x 10^7	104.67	104.77	104.56	104.43	104.43	0.036
Mínimo	0.00	0.04	0.051	-0.09	0.049	0.049	-0.84
Máximo	1.98 x 10^9	1356.69	1356.69	1356.69	1356.69	1356.69	1.90

Tabela 11 – Estatísticas descritivas do conjunto de dados filtrado referente ao ano de 2016.

2017	Volume	Close	High	Adj Close	Low	Open	Returns
Média	2.62 x 10^6	34.62	34.87	33.05	34.35	34.62	0.001
D. P.	1.37 x 10^7	135.52	137.72	135.50	135.38	135.59	0.039
Mínimo	0.00	0.13	0.13	-0.18	0.12	0.13	-5.146
Máximo	8.73 x 10^8	1966.80	1966.80	1966.80	1966.80	1966.80	1.652

Tabela 12 – Estatísticas descritivas do conjunto de dados filtrado referente ao ano de 2017.

2018	Volume	Close	High	Adj Close	Low	Open	Returns
Média	2.99 x 10^6	45.62	45.99	44.41	45.26	45.64	0.0006
D. P.	1.24 x 10^7	252.76	253.64	252.75	252.05	252.79	0.028
Mínimo	0.00	0.19	0.19	0.19	0.18	0.19	-0.44
Máximo	1.38 x 10^9	4255.35	4329.20	4255.35	4238.12	4238.12	1.487

Tabela 13 – Estatísticas descritivas do conjunto de dados filtrado referente ao ano de 2018.

2019	Volume	Close	High	Adj Close	Low	Open	Returns
Média	4.34 x 10^6	55.81	56.15	55.04	55.41	55.80	0.002
D. P.	2.26 x 10^7	290.70	291.01	290.69	290.36	290.68	0.025
Mínimo	0.00	0.19	0.20	0.19	0.19	0.19	-0.35
Máximo	1.90 x 10^9	3879.43	3903.44	3879.43	3879.43	3903.44	0.93

Tabela 14 – Estatísticas descritivas do conjunto de dados filtrado referente ao ano de 2019.

Dos resultados acima, vê-se a evolução no preço médio de fechamento ajustado ("Adj Close", variável utilizada para cálculo do retorno de cada ação) a partir do ano de 2016, posterior a recessão que o país enfrentou nos anos anteriores, movimento que também pode ser visto de maneira similar no gráfico do IBOVESPA, Figura 10. Os consideráveis prejuízos (registrados pelas mínimas no campo "Returns") acima de 32% saltam aos olhos do investidor aumentando a motivação na composição da carteira mais rentável acerca do risco, objetivo motivador deste projeto.

6.4.2 Divisão dos Dados

Não existe uma proporção ótima para divisão de conjunto de dados. Segundo especialistas, a abordagem mais comum é a divisão dos dados em 67% dedicados a etapa de treinamento e 33% para as outras etapas, (BROWNLEE, 2020) e (PALANISAMY, 2006). A Tabela 15 demonstra a divisão do conjunto de dados utilizados neste trabalho.

Conjunto de Dados	**Número de Linhas**	**Data Inicial**	**Data Final**
Treino	187844 (67%)	2012-06-19	2017-06-28
Validação	47112 (16.5%)	2016-06-29	2018-09-20
Teste	47112 (16.5%)	2018-09-21	2019-12-30

Tabela 15 – Composição dos Conjuntos de Treino, Validação e Teste

Em Treino, utiliza-se os dados para a construção das carteiras e, no caso da metodologia por inteligência artificial, construção do modelo a ser utilizado. Em Validação, o objetivo é simular os resultados de cada carteira de investimento nos primeiros meses após o intervalo de tempo utilizado para sua construção. A partir destes resultados, calcula-se as projeções da rentabilidade de cada carteira para o período do conjunto de dados de teste que fornecem ao investidor informações relevantes para a escolha da carteira ideal a ser utilizada no próximo período (Conjunto de Dados de Teste - simulando a aplicação da carteira de investimentos e verificar sua eficácia). Neste projeto todas as carteiras geradas foram aplicadas ao período de Teste para fins comparativos. Os resultados das projeções do conjunto de validação podem ser vistos no Apêndice A.

6.5 Experimentos - Markowitz

Para os experimentos com Markowitz, gerou-se uma coleção de carteira de investimentos e realizou-se o cálculo de suas Expectativas de Retorno e Risco,R_{cart} e σ_{cart}, com o intuito de construir a Fronteira Eficiente para seleção da carteira GMV.

Para o desenvolvimento do conjunto de dados para a análise por Markowitz, uma abordagem computacional foi empregada, visando facilitar a construção da análise de Markowitz e otimizá-la, e através das ferramentas Python e Numpy gerou-se uma coleção de 1000 carteiras de investimentos em que a distribuição do peso de cada ativo foi definido de maneira aleatória. Cada carteira teve seus pesos normalizado entre 0 e 1, através do processo de divisão pela soma de todos os valores. A Figura 13 demonstra a constituição de um amostra de 10 carteiras e o peso de cada ativo.

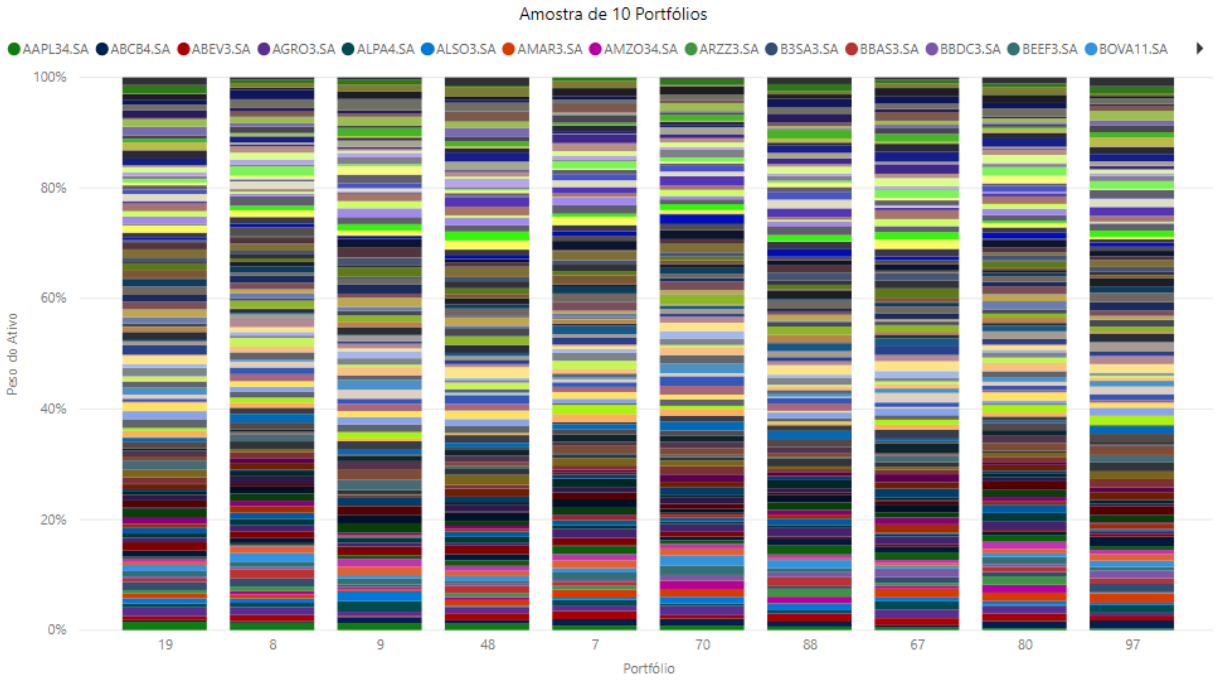


Figura 13 – Distribuição do Peso dos Ativos em uma amostra de 10 Portfólios do conjunto gerado

As 1000 carteiras foram implementadas dentro dos dados de teste em que foram verificados o seu retorno médio e volatilidade, anualizando-os, equações 2.4 e 2.5, para termos de 252 dias. A carteira de GMV foi selecionada a partir do portfólio de menor variância presente na tabela performance.

A carteira GMV foi implantada para simulações dentro intervalo de tempo para o conjunto de dados de Validação e Teste.

6.6 Experimentos - Critério de Kelly

Como parâmetro para o cálculo do Critério de Kelly, as probabilidades de cada ativo vencer e perder precisam ser calculadas. Desta forma, para calcular tal probabilidade foi calculado a partir dos dados de treino o número de vezes em que cada ativo obtinha retorno positivo, ou seja, $R_{acao} > 0$, e dividiu-se pelo número de pregões. Da teoria básica de probabilidades, pôde se extrair a probabilidade de uma ação perder a partir de $Pr(perder) = 1 - Pr(vencer)$.

Para verificar diversos padrões e como a probabilidade impacta na construção das carteiras, construiu-se quatro portfólios o qual o número de dias anteriores foram:

- 1244 dias: Analisando-se todo o conjunto de treino;
- 500 dias: Analisando os dados de Junho de 2015 a Junho de 2017;

- 150 dias: Analisando os dados de Novembro de 2016 a Junho de 2017;
- 60 dias: Analisando os dados de Abril de 2017 a Junho de 2017;

Após o cálculo de cada fração a ser alocada, a carteira foi normalizada entre o intervalo 0 e 1. Não houveram outras modificações no conjunto de dados para a aplicação dos experimentos de Kelly.

6.7 Experimentos - Aprendizado por Reforço

O problema da construção de carteira de investimentos no mercado de ações pode ser interpretado como um POMDP - Processo de Decisão de Markov Parcialmente Observável - por não ser possível obter informações sobre todos os estados do ambiente, como por exemplo, o comportamento de outros investidores, tomadas de decisão dos executivos de cada empresa ou ações regulamentares do governo.

O conjunto de estados foi obtido pela fração do preço de fechamento (*Close*, Tabela 3) pelo preço de abertura (*Open*, Tabela 3) do dia atual e seus dois dias anteriores, calculado para cada ativo presente nos dados. Desta forma, os estados foram compostos por 453 dimensões (Número de Ações x 3) como pode ser visto a definição de s_t abaixo.

$$s_t = [\frac{c_{0,0,t}}{o_{0,0,t}}, \frac{c_{0,1,t}}{o_{0,1,t}}, \frac{c_{0,2,t}}{o_{0,2,t}}, ..., \frac{c_{N,0,t}}{o_{N,0,t}}, \frac{c_{N,1,t}}{o_{N,1,t}}, \frac{c_{0,2,t}}{o_{0,2,t}}] \tag{6.1}$$

Em que:

- s_t é o estado no tempo t;
- $c_{n,w,t}$ é o valor de fechamento do ativo n no dia da janela w (janela de três dias anteriores ao dia t) no tempo t;
- $o_{n,w,t}$ é o valor de abertura do ativo n no dia da janela w (janela de três dias anteriores ao dia t) no tempo t;

A ação do agente quando submetido ao estado s_t é a distribuição dos pesos dos ativos para a carteira de investimentos, em que cada peso está dentro do intervalo [0,1] e a somatória de todos os pesos deve resultar em 1. Observação: A carteira de investimentos sugerida pelo modelo deve ser utilizada na simulação do próximo dia (*t+1*), pois o estado s_t foi construído a partir dos resultados do ativo no dia t.

$$a_t = [\omega_{0,t+1}, ..., \omega_{N,t+1}], \sum_{i}^{N} \omega_i = 1 \tag{6.2}$$

Em que:

- a_t é a ação executada pelo agente no estado s_t
- $\omega_{n,t+1}$ é o peso do ativo proposto pelo modelo no dia t para ser simulado no dia *t+1*.

A recompensa R_t que o agente recebe no dia t é o retorno financeiro da carteira de investimentos sugerida no dia anterior *t-1*.

$$R_t = \sum_{i}^{N} r_{i,t}\omega_{i,t-1} \tag{6.3}$$

A Tabela 6.7 expõe a dimensão da tabela final de treinamento para o modelo de Aprendizado por Reforço

Conjunto de Dados	**Dimensão**
Observações	451 Colunas
Ações	Vetor de 151 elementos

Tabela 16 – Tabela de Dimensões do conjunto de dados Observações e Ações

Similar a ideia de (CONEGUNDES; PEREIRA, 2020) e (H. et al., 2020), o espaço de observações foi composto por dados de dias anteriores ao dia de treinamento t, o primeiro utilizando valores da proporção de $\frac{Close}{Open}$ apenas e o segundo valores extraídos de técnicas de movimentação de mercado como MACD. O resultado de Conegundes implicou na melhor eficiência no reforço financeiro utilizando-se uma janela de três dias.

Como comentado anteriormente, o ambiente de treinamento foi construído através da ferramenta da Open AI - Gym (BROCKMAN et al., 2016) que tem como proposta a facilidade e versatilidade para execução de um modelo único e replicável e diversidade dos dados. Foi desenvolvida uma lógica dentro do ambiente de treinamento do Gym para que a cada novo episódio o conjunto de dados de treino fosse repartido em intervalos de tempo aleatórios de no mínimo 30 e máximo de 444 dias, com o objetivo de fazer que o agente experienciasse o máximo de diversas situações de comportamento do mercado. O modelo foi treinado sob 250000 passos de tempo, aproximadamente 1000 episódios, utilizando os algoritmos DDPG e PPO por recorte para treinamento. Planejando uma abordagem em que os modelos estejam mais capacitados para trabalharem com dados mais próximos ao fim do conjunto de treino, após 700 episódios era definido para que o treinamento acontecesse apenas dentro do intervalo dos últimos 400 dias.

6.7.1 Configuração das Redes Neurais Ator-Crítica

Como comentado anteriormente, PPO e DDPG são métodos de Aprendizado por Reforço Profundo que utilizam de rede neural Ator-Crítica para modelagem de suas

políticas. Abaixo, demonstra-se as configurações das redes neurais artificiais referentes a sua classe, estrutura das camadas e função de ativação para execução dos experimentos.

- **PPO:**
  - **Classe:** *Multi-Layer Perceptron*;
  - **Camadas:** 2 camadas compostas por 64 nós cada;
  - **Função de ativação:** Tangente hiperbólica (tanh);
- **DDPG:**
  - **Classe:** *Multi-Layer Perceptron*;
  - **Camadas:** 2 camadas compostas por 64 nós cada;
  - **Função de ativação:** Unidade Linear Retificada (ReLU);

A Figura 14 ilustra o diagrama das redes neurais Ator-Crítica utilizada pelos modelos, em que E é o número de nós na camada de Entrada e S o número de nós na camada de Saída. Para a rede neural Ator, em que o sinal de entrada é o estado atual do agente e a saída é sua ação, E = 403 e S = 151. Para a rede neural crítica, em que o sinal de entrada é o estado atual mais a ação sugerida pela rede ator e o sinal de saída é a possível recompensa a longo prazo, E = 554 e S = 1.

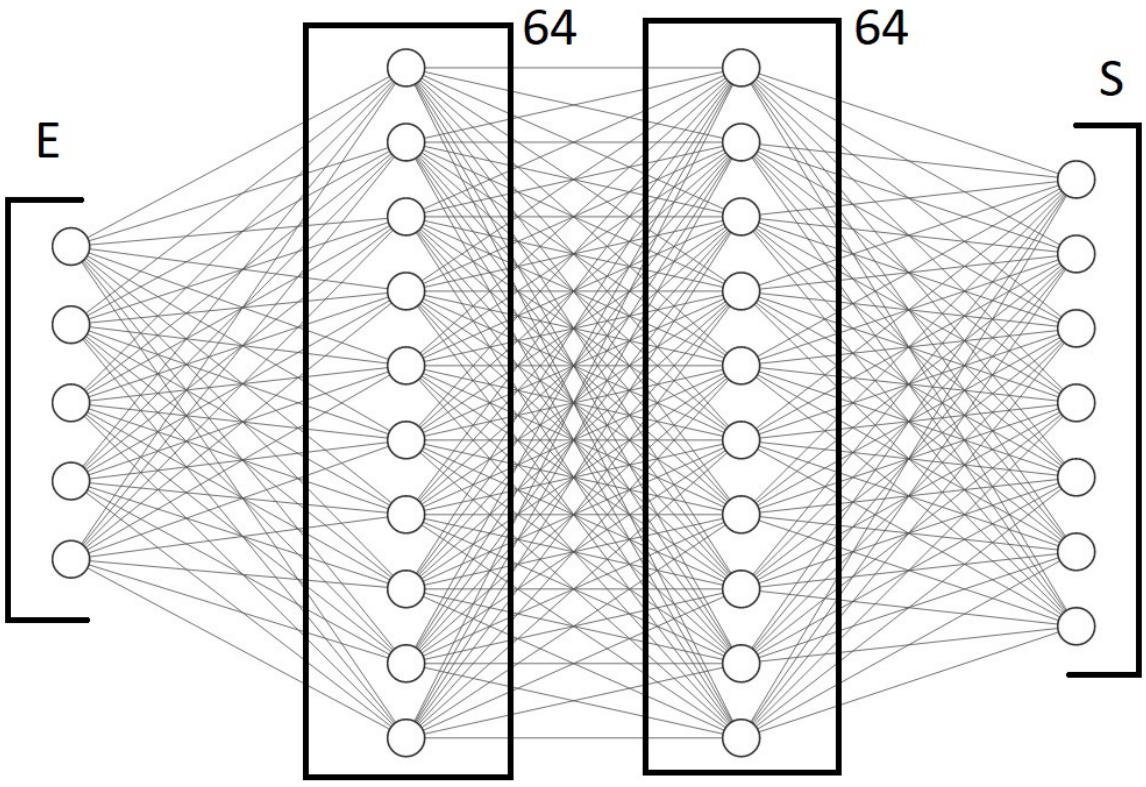


Figura 14 – Diagrama das redes neurais Ator-Crítica

7 Resultados e Discussão

Os resultados obtidos nas execuções dos experimentos foram analisados considerando a particularidade de cada estratégia e a tendência de curto e médio prazo considerando os períodos de validação e teste. A carteira construída pelo Modelo de Markowitz foi satisfatória e constituiu-se o terceiro portfólio mais rentável após 620 dias, quando atingiu retorno de aproximadamente 121% e risco associado às operações diárias de 0.94%. Como esperado de uma carteira gerada por uma metodologia de sistema fechado, viu-se que os resultados acompanhavam a tendência do período em que a estratégia foi implementada (período de treinamento). Este comportamento pode ser visto logo na comparação com os resultados do período de Validação, sendo a carteira menos rentável dentre todas as metodologias e IBOVESPA, ficando acima apenas da inflação. Seu desempenho computacional durante o período de treinamento foi o que obteve o tempo de processamento mais custoso, devido ao esforço operacional repetitivo de se calcular os retornos diários para cada carteira gerada. Partindo do princípio de que mais carteiras pudessem alcançar melhores resultados, o custo da aplicação da metodologia cresceria (consideravelmente). Para Kelly, foram construídas 4 carteiras seguindo aplicando-se a fórmula com probabilidades contabilizadas para 60, 150, 500 e 1244 dias. Foi visto que para as carteiras de intervalos mais longos, menor se seguiam as tendências de curto prazo, apresentando menores variações máximas (queda máxima) que as carteiras de 30 e 150 dias. Por outro lado, as carteiras de maiores intervalos acabam perdendo oportunidades repentinas pela maior distribuição dos pesos em seu catálogo. A abordagem por aprendizado por reforço obteve os melhores resultados para as duas carteiras construídas, com 152.6% de acúmulos em DDPG e 151.3% em PPO. O DDPG manteve dominância constante sobre as outras carteiras e os benchmarks durante todo o período de validação e teste. Ainda, foi a única carteira que superou o IBOVESPA durante a etapa de validação.

Este capítulo está organizado para inicialmente cobrir e expor os resultados de cada metodologia para, ao final, concentrado na seção de Aprendizado por Reforço, analisar comparativamente todas as carteiras e discutir o desempenho computacional.

7.1 Carteiras de Markowitz

Como requisito à aplicação das Regras de Markowitz e construção da Froteira Eficiente, é necessário a obtenção de uma coleção de portfólios de ativos junto de suas métricas relativas a **Expectativa de Retorno** e **Volatilidade**. Portanto, de maneira pseudo-aleatória, foram gerados 1000 portfólios de ativos e no passo seguinte foram calculados suas médias e volatilidade anualizadas, como pode ser visto na Figura 13 em

Métodos.

A fim de apoiar visualmente a análise de seleção de portfólios e determinar a Fronteira Eficiente, o Gráfico de Dispersão da Figura 15 foi construída utilizando-se dos cálculos da Expectativa de Retorno Financeiro e Volatilidade de cada carteira. Para a escolha da carteira de Markowitz, selecionou-se a carteira de menor volatilidade, chamada de carteira GMV. A carteira 73 ($R = 12.24\%$ e $\sigma = 15.16\%$) foi a escolhida e está destacada na Figura 15.

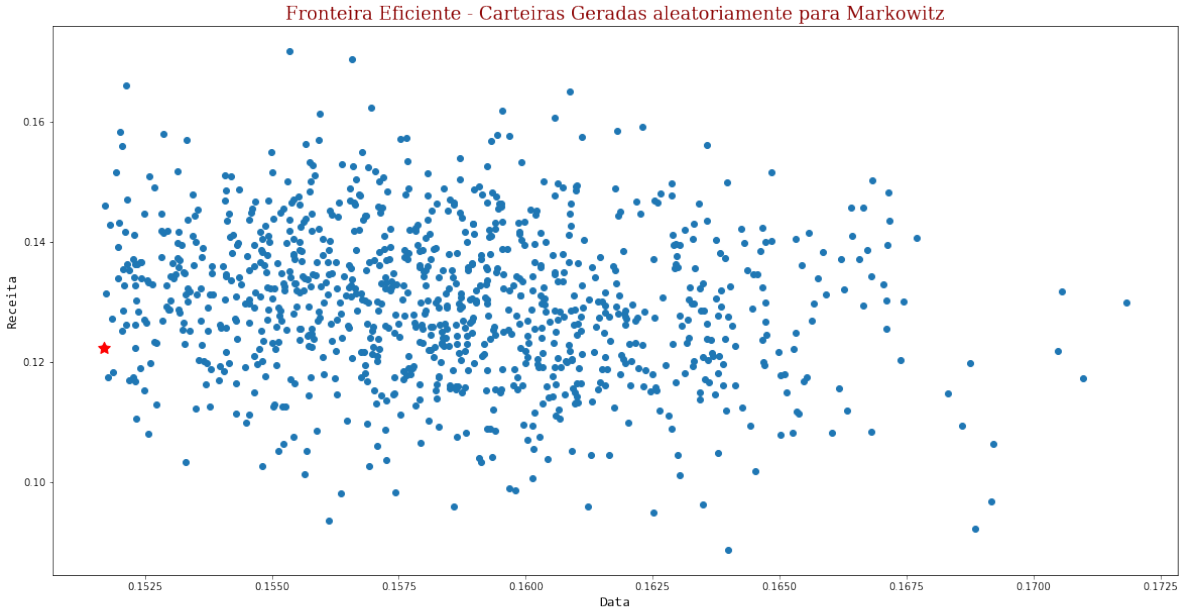


Figura 15 – Gráfico de Dispersão dos Retornos Financeiros e Volatilidade anualizados dos portfólios

O Conjunto de Dados de Validação foi constituído dos dados históricos relativos ao período de Junho de 2017 a Setembro de 2018, compondo aproximadamente 13,6% da base total histórica utilizada para avaliação. O objetivo desta base de dados é obter-se uma expectativa do comportamento do portfólio a um período de dados anterior aos testes. A Figura 16 demonstra o comportamento da carteira de Markowitz durante o período de Validação.

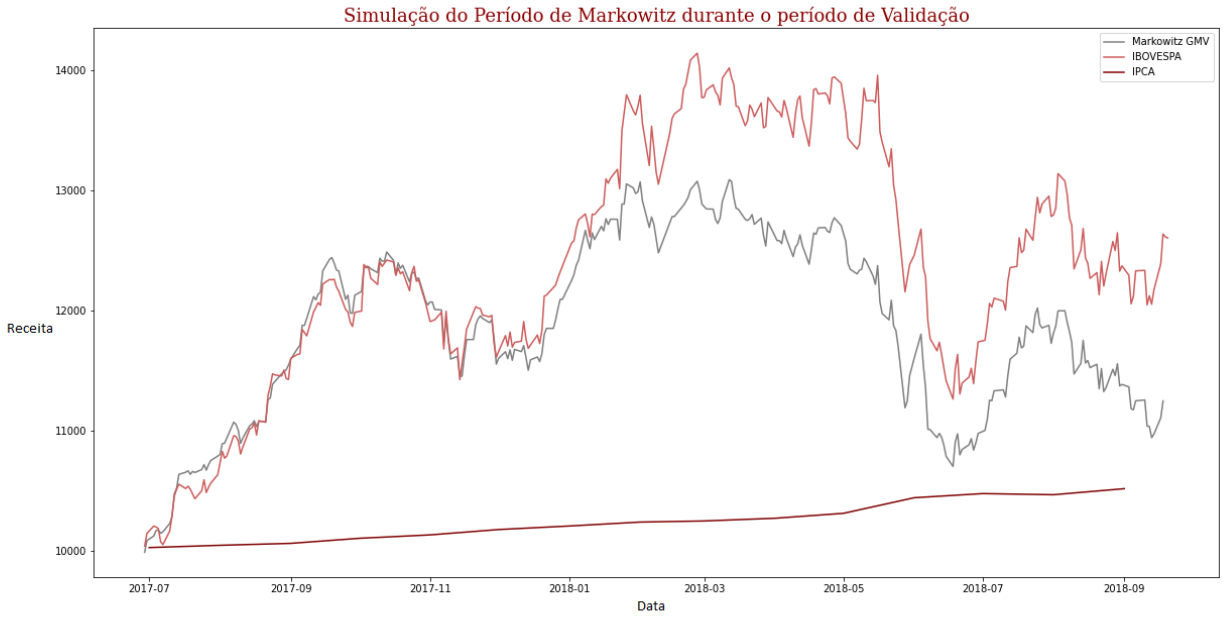


Figura 16 – Simulação da Carteira de Markowitz durante o Período de Validação

A Tabela 7.1 mostra a Média de Retorno Esperado e Média da Volatilidade para os próximos 312 dias, baseado nos resultados obtidos no período de validação.

Carteira	**M.R.E (312)**	**M.V. (312)**
Markowitz	50,8%	16,7%

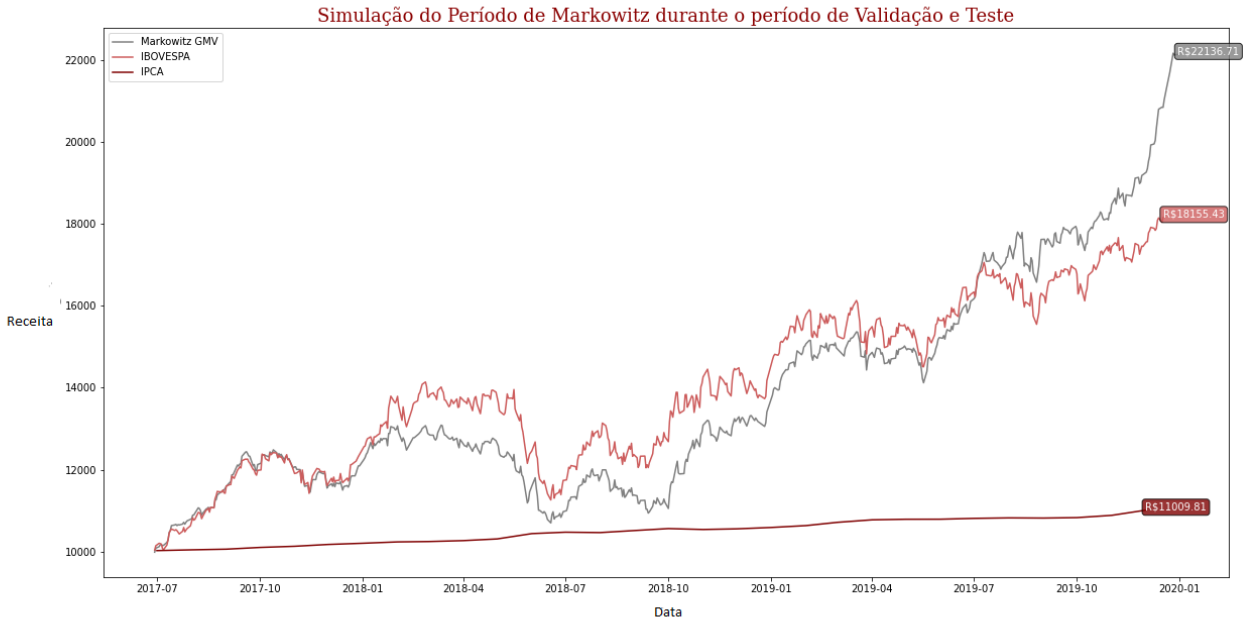


Figura 17 – Simulação da carteira de Markowitz no período de Validação e Teste

Por outro lado, a carteira de Markowitz teve alto rendimento quando aplicado para em conjunto com o período de teste, como pode ser visto na 17, superando as marcas de projeção 24. Elevando o aporte a R$22136,17 reais, o investimento aumentou ⬆121.36%, em comparação aos 81,5% do mercado e 10% da inflação.

7.2 Carteiras por Critério de Kelly

O Modelo de Markowitz (Modelo Média-Variância), um dos pioneiros para a construção da Teoria Moderna de Portfólios, vem desde o século 20 sendo muito estudado pela academia que, porém, vem também acumulando série de críticas devido suas hipóteses pré estabelecidas sobre o investidor e seu ambiente não se adequarem ao mundo real (MANGRAM, 2013).

7.3 Simulação no Conjunto de Dados de Validação

Para que os portfólios de Kelly fosse obtido, foi analisado a quantidade de dias anteriores para se considerar no cálculo da probabilidade e dos ganhos recebidos. Alguns experimentos foram realizados variando-se o intervalo do número de dias analisados, executando experimentos para 1244, 500, 100 e 60 dias, e verificando o comportamento no intervalo de Validação.

A Figura 18 demonstra o comportamento resultante da análise:

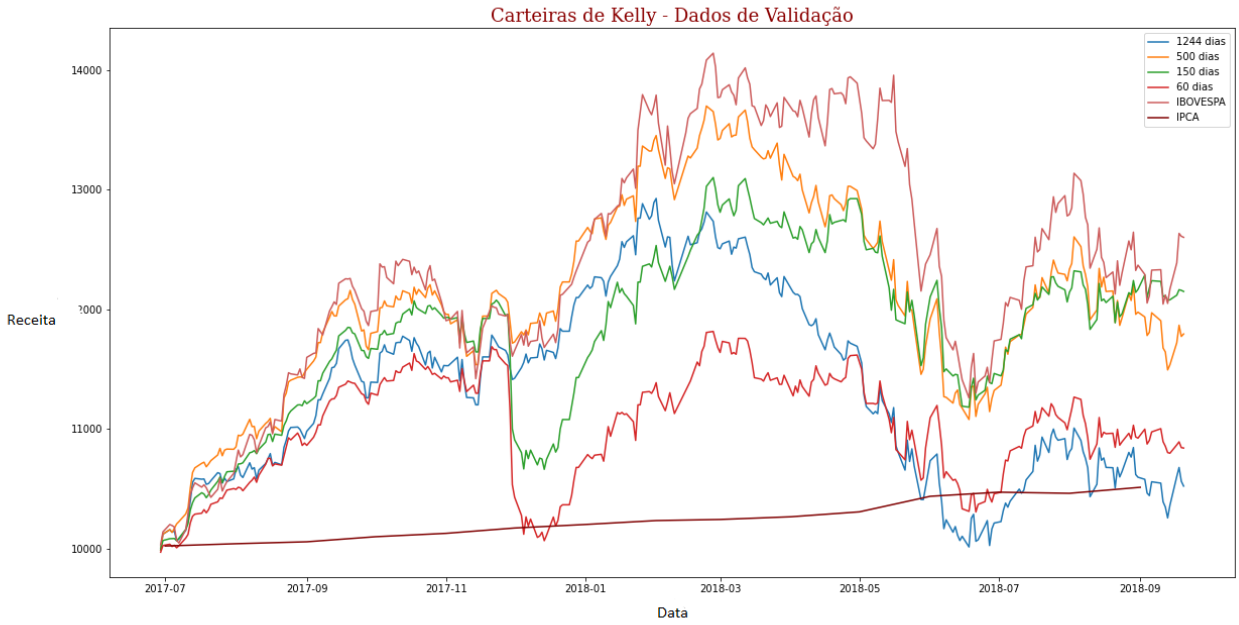


Figura 18 – Comportamento do Critério de Kelly para 60, 100, 500 e 1244 dias

Para o conjunto de Validação, vê-se que os portfólios gerados por Kelly deram resultados razoavelmente satisfatórios. A receita mostrou aumento em todas carteiras ao final da simulação e ficou acima da valorização da inflação; não obstante, todas as carteiras tiveram resultados inexpressivos frente a valorização do mercado relativa ao IBOVESPA. A Tabela 17 demonstra a performance das carteiras de Kelly durante o período de teste.

Carteira	**Retorno (R)**	**% IBOVESPA**	**% IPCA**
1244 dias	⬆5,24%	⬇-16%	⬆0,09%
500 dias	⬆17,9%	⬇-6,4%	⬆12,18%
150 dias	⬆21,5%	⬇-3,5%	⬆15,55%
60 dias	⬆8,4%	⬇-13,9%	⬆3,1%

Tabela 17 – Performance das carteiras do Critério de Kelly para o período de Validação

Porém, quando avalia-se os resultados durante todo o período (Validação + Teste), como pode ser visto na Figura 19, vê-se que as carteiras de Kelly superam o mercado e atingem resultados expressivos para as carteiras de 150 e 500 dias, com a carteira de 150 dias com receita similar ao modelo de Markowitz.

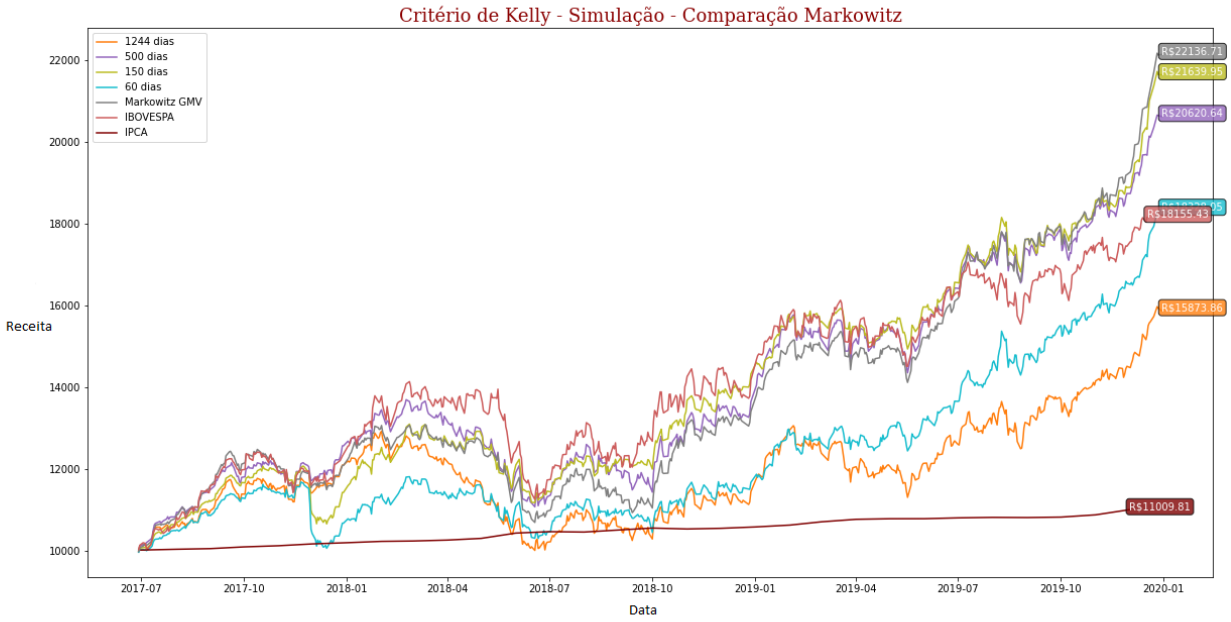


Figura 19 – Simulação Carteiras de Kelly no Conjunto de Dados Validação e Teste

Carteira	**Queda Máxima**
Kelly 1244	-3,1%
Kelly 500	-3,3%
Kelly 150	-5,2%
Kelly 60	-5,5%

Tabela 18 – Tabela com as quedas máximas das carteira de Kelly

A Tabela 18 mostra o valor referente ao dia em que mais se obteve um retorno negativo. Vê-se que a carteira referente a 60 dias foi a carteira que obteve a maior queda enquanto a carteira de Kelly 1244 dias obteve a menor. Este comportamento é esperado de uma abordagem probabilística o qual a probabilidade é contabilizada pelos dias anteriores. Nas carteiras o qual se teve mais dias analisados, as probabilidades, e, consequentemente a distribuição dos pesos entre os ativos estarão mais distribuídos evitando que resultados pontuais e foras do padrão impactem negativamente no rendimento da carteira (mês de Maio de 2018 é um exemplo, o qual a queda na carteira de 1244 dias foi bem pequena e a de 30 dias apresenta um maior degrau). Por outro lado, este perfil de carteira irá arriscar menos, deixando de aproveitar os rendimento.

7.4 Carteira por Aprendizado por Reforço

As carteiras de investimentos geradas pelo modelo de Aprendizado por Reforço tiveram resultados satisfatórios, revelando o melhor desempenho dentre as metodologias expostas no presente trabalho. O referido modelo obteve 151% de valorização total do aporte em simulação entre o período de Validação e Teste, com volatilidade média de 0.9% nos retornos diários.

Durante o treinamento do modelo, foi coletado a média de retornos acumulados obtidos para verificar o comportamento do aprendizado. A Figura 20 demonstra o comportamento para o treinamento do algoritmo por PPO e a Figura 21 para a DDPG. Vê-se que as recompensas antes do episódio 700 é disperso e com alta variância. Isto ocorreu devido a modelagem da fase de treinamento definida em que a partir do episódio 700, o ambiente do Gym disponibilzava apenas dados referentes aos últimos 400 dias dos dados de treino, com o âmbito de apontar ao agente os dados mais recentes aos próximos que serão simulados.

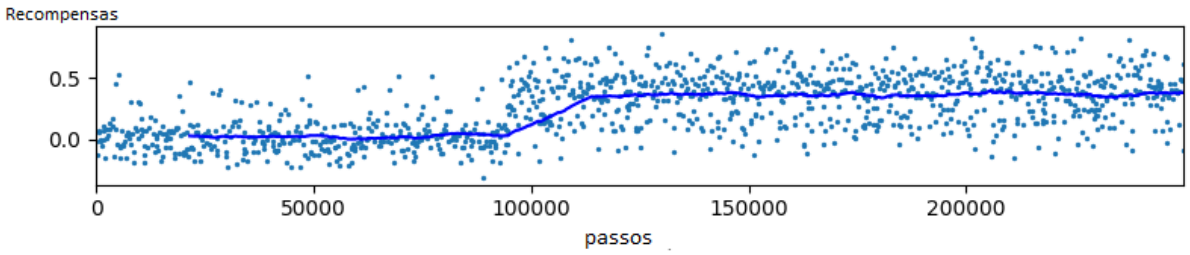


Figura 20 – Análise das Recompensas durante o período de Treino para o modelo de PPO

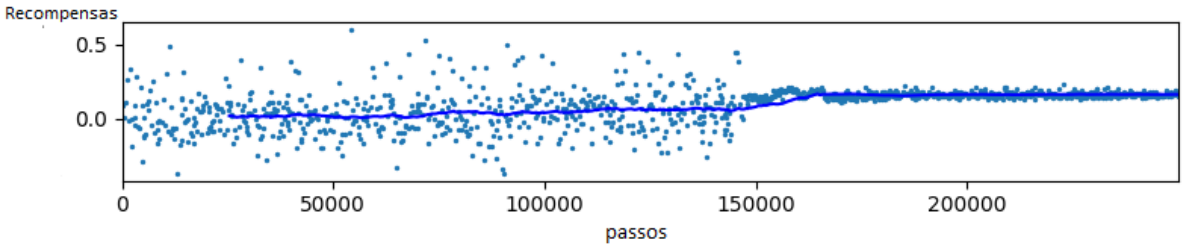


Figura 21 – Análise das Recompensas durante o período de Treino para o modelo de DDPG

Durante a etapa de validação, foi visto que, assim como nas carteiras produzidas por Markowitz e Kelly, o modelo por PPO apresentou resultados positivos, com performance que pode ser vista na Tabela 19, porém ficou levemente abaixo do mercado mas acima da

inflação. Já o modelo desenvolvido por DDPG atingiu resultados acima de todas as outras estratégias e fechou o período com alta de rotatebox90➡31.58%, isto é, 4.41% quando comparado ao IBOVESPA.

A Figura 22 demonstra o comportamento das carteiras por aprendizado por reforço durante o período de validação.

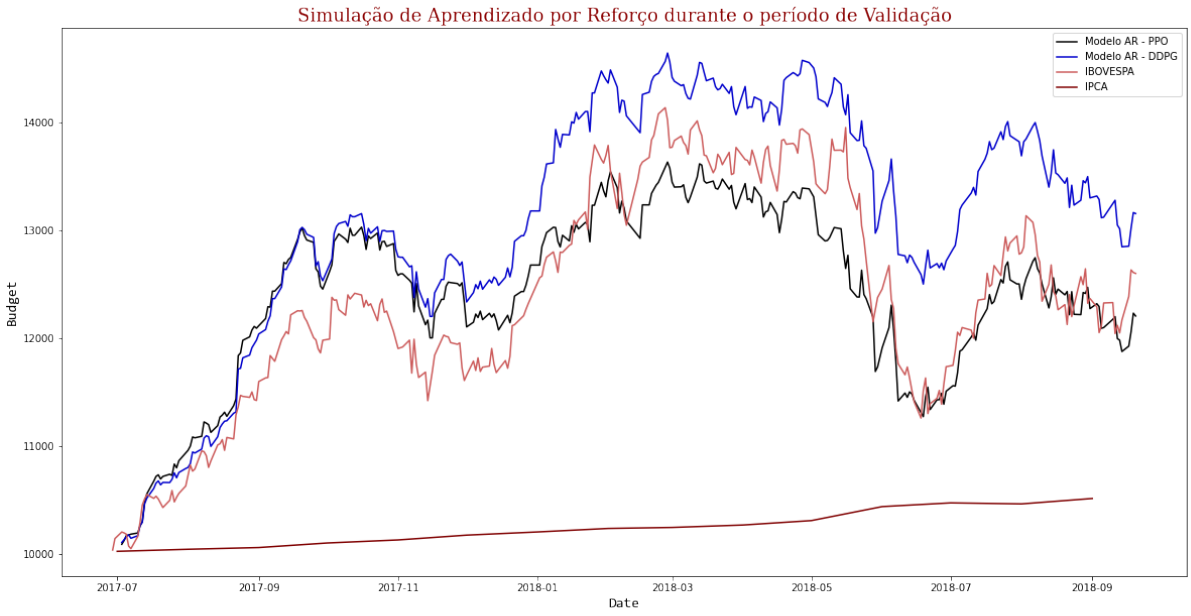


Figura 22 – Simulação no Conjunto de Validação para a carteira gerada por Aprendizado por Reforço

Porém, como visto na Figura 22, o modelo de Aprendizado por Reforço DDPG e PPO foram os únicos que ficaram acima do IBOVESPA no intervalo entre Julho de 2017 e Janeiro de 2018, diferente das carteiras expostas em momento prévio.

Quando olha-se a perspectiva a longo-prazo, vemos que as carteiras são as que obtém os principais retornos e a que demonstra retornos mais consistentes, como pode ser visto na Figura 23 e na Tabela 19.

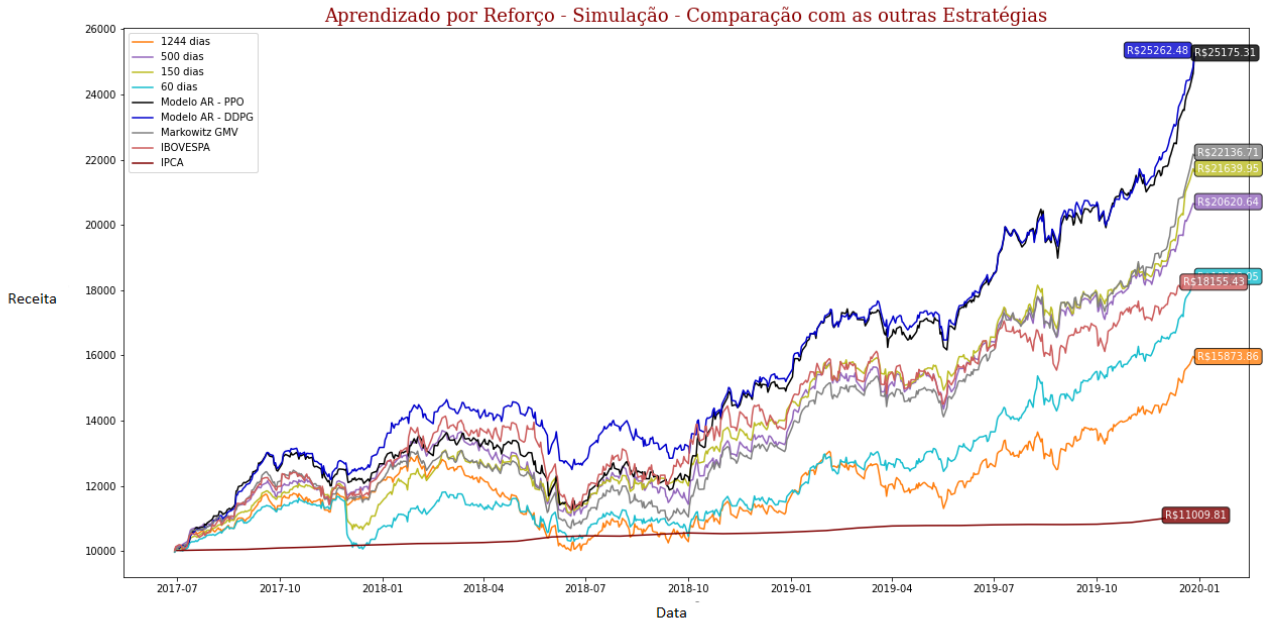


Figura 23 – Comparação da carteira por Aprendizado por Reforço com as outras técnicas

Carteira	Retorno Acumulado	Volatilidade
PPO	151,7%	0,98%
DDPG	152,6%	0,93%
Markowitz	121,3%	0,94%
Kelly 1244	58,7%	1,02%
Kelly 500	106,2%	1,04%
Kelly 150	116,3%	0,91%
Kelly 60	83,3%	0,91%

Tabela 19 – Tabela de Performance - Período de Validação e Teste - Todas as Simulações

Da Tabela 19 vê-se que as carteiras construídas por Inteligência Artificial ficaram, respectivamente, 30.4% e 31.3% acima da segunda colocada, Markovitz, e 35% acima da carteira mais rentável, gerada por Kelly. Além disso, pode ser visto que a carteira superou as expectativas levantadas pelas projeções 26 e que durante toda jornada a carteira aparece como a mais rentável frente a todas as outras, com o cenário atípico da PPO que ficou atrás ao *benchmark* de mercado durante o intervalo de Janeiro a Junho de 2018. Ainda na Tabela 19, nota-se a volatilidade da carteira de DDPG que, apesar de sua maior rentabilidade, foi a metodologia com menor risco frente aos 4 principais portfólios.

Carteira	Queda Máxima
A.R. PPO	-4,1%
A.R. DDPG	-4,2%
Markowitz	-4,3%
Kelly 1244	-3,1%
Kelly 500	-3,3%
Kelly 150	-5,2%
Kelly 60	-5,5%

Tabela 20 – Tabela com as quedas máximas de cada carteira

Da Tabela 20 podemos ver que as carteira de Aprendizado por Reforço por PPO, durante todo o intervalo de Validação e Teste, apresenta uma queda máxima de apenas 4.1%, acontecendo no dia 29 de Maio de 2018, no meio do período em que todas as carteiras ficaram abaixo do IBOVESPA, e ainda assim melhor que o segundo colocado em 0.2%.

7.5 Desempenho Computacional

Para a análise de desempenho computacional, buscou-se medir o impacto do emprego da computação o tempo de processamento para conclusão do experimento e a quantidade de alocação de memória, mais específico ao problema do Aprendizado por Reforço.

A Tabela 21 demonstra o tempo, em segundos, gasto para a execução de experimentos para desenvolvimento de uma carteira.

Carteira	Tempo de Processamento
Markowitz	4929,19 segundos
Kelly	12,3 segundos
PPO	379,28 segundos
DDPG	1224,54 segundos

Tabela 21 – Tabela de Tempo de Processamento para a execução do experimento

Da Tabela 21 pôde se ver que a metodologia de Markowitz foi a mais custosa em questão de tempo de processamento, em seguinda a abordagem por Aprendizado por Reforço através do modelo de DDPG. Isto ocorreu devido a necessidade de se percorrer todo o Conjunto de Treino (1244 linhas) e calcular as métricas de performance proporcionalmente ao número de carteiras aleatoriamente geradas, no caso 1000. Já a abordagem de Aprendizado por Reforço por DDPG necessita de um esforço computacional considerável devido a computação dos pesos das Redes Neurais Profundas que utiliza internamente atualização dos pesos dos gradientes para a computação do treinamento, já na modalidade PPO, que possui uma implementação mais simplificada, teve um tempo de processamento

mais moderado. Além do mais, para ambas abordagens, conforme se aumentasse seus parâmetros (número de carteiras aleatoriamente geradas por Markowitz e número de Episódios analisados pelo aprendizado) o tempo de processamento para a construção da carteira também evoluia. Por outro lado, Kelly pôde calcular a distribuição da carteira simplesmente iterando uma vez sobre o conjunto de treino e aplicando a fórmula de Kelly, algo computacionalmente pouco custoso.

Diferente das outras abordagens, o modelo de Aprendizado por Reforço precisa de uma fase de treinamento para que possa ser executado em simulações. A etapa de treinamento, como explicitado em capítulos anteriores, consiste na aplicação do algoritmo alvo que irá realizar o reconhecimento dos padrões a partir de um conjunto de dados de treinamento.

Para o treinamento do modelo selecionado, como mencionado na seção referente 650 episódios analisados pelo modelo, resultando em um tempo de processamento de 12 minutos na máquina alvo (707 segundos).

Como o modelo de aprendizado por reforço precisou de uma construção de um conjunto de dados específico para consumo, consistindo na modelagem do Espaço de Estados e Ações, isto fez com que esta metodologia utilizasse aproximadamente 33% a mais no armazenamento. Como mostrado ainda a etapa de métodos, as tabelas modeladas para os espaços de observação tem crescimento colunar em W vezes o número de ativos analisados pela estratégia, em que W é a janela de dias anteriores ao dia analisado. Para análises mais complexas utilizando aprendizado por reforço, utilizando-se de W como dados mensais ou anuais, a memória será uma variável pertinente durante a fase de planejamento.

8 Conclusão

Este projeto de Trabalho de Conclusão de Curso consistiu no estudo de métodos de desenvolvimento de carteiras de investimento, seguindo a abordagem em sistema fechado (Markowitz), probabilística (Kelly) e por inteligência artificial (Aprendizado por Reforço). Após a implementação dos experimentos, foi executada uma análise comparativa dos rendimentos financeiros de cada carteira em um conjunto de dados históricos de 151 ativos da bolsa de valores de São Paulo, entre os anos de 2012 a 2019.

Dos resultados presentes no capítulo anterior, concluiu-se que é válido utilizar o Aprendizado por Reforço para elaborar carteiras de investimento. As metodologias DDPG e PPO obtiveram rendimentos maiores que 150% sendo a primeira o modelo menos volátil entre as cinco principais carteiras (Tabela 19). Além disso, o modelo de Aprendizado por Reforço de DDPG dominou o resultado acumulado na maioria dos pregões simulados, visto que é uma metodologia dinâmica e pode sugerir novas carteiras de investimentos para cada dia e o modelo por PPO ficou atrás do principal índice do mercado o IBOVESPA por um curto período de 4 meses em 2018 (Figura 22).

O modelo de Markowitz, mesmo que de sistema fechado, apresentou resultados superiores a abordagem probabilística mas que pecou no desempenho computacional, sendo aproximadamente 4 vezes mais custosa de processamento que a abordagem por Aprendizado por Reforço (Tabela 21). O critério de Kelly mesmo sendo a estratégia menos rentável frente aos outros métodos, obteve resultados satisfatórios para a rentabilidade em sua carteira de análise de 150 dias passados. Das carteiras do critério de Kelly pode se notar que para carteiras de maior intervalo de dias analisados, a chance de seguir uma tendência a curto prazo é menor, resultando em um valor de queda máximo reduzido em frente às outras carteiras (Tabela 20) mas que por outro lado, evita que este tipo de carteira se aproveite de retornos positivos momentâneos.

Muitos passos podem ser explorados que podem ser explorados a partir deste trabalho, como a utilização de outros métodos de Aprendizado por Reforço como SAC (HAARNOJA et al., 2018) ou A2C ((MNIH et al., 2016)). Neste trabalho, o treinamento do modelo de aprendizado concluía após a execução do número de episódios previamente definido, sendo esta uma limitação. São válidos para trabalhos futuros abordagens que definam critérios de parada explícitos durante o treinamento, almejando o encontro dos pontos máximos da função recompensa. Outra aplicação para estudos é a utilização de outras variáveis das ações da bolsa de valores brasileira. (H. et al., 2020) propôs um robusto modelo para o espaço de estados, composto de métricas da análise técnica de movimentação de investimentos como MACD (Média Móvel Convergente e Divergente) e RSI (Índice

de Força Relativa). Outra abordagem quanto a composição do conjunto de estados é a utilização de dados referentes a fatos macroeconômicos brasileiro, tal qual podem ser estimados econometricamente. Dessa forma, o modelo de aprendizado por reforço poderia aumentar sua adaptabilidade e diminuir seu risco em cenários de quedas significativas da bolsa, como foi visto nos períodos de recessão em 2008 ou 2014-16. Seguindo esta linha, a utilização do PIB como variável pode favorecer o aumento na rentabilidade por sinalizar a entrada de capital no país por investidores estrangeiros.

Como citado no capítulo de Revisão Teórica, (MACLEAN; ED, 2006) cita que para abordar o mercado financeiro utilizando-se do critério de Kelly é mais viável utilizar-se de uma abordagem contínua dos dados e calcular as distribuições de probabilidade dos ativos, visando obter melhores resultado e aplicação do critério de Kelly contínuo aos dados da bolsa de valores brasileira pode ser um novo caminho para a obtenção de resultados superiores a abordagem discreta presente neste trabalho.

Por último, os dados utilizados neste trabalho consistiam em dados históricos até 2019. Em 2020, a pandemia da COVID-19 demonstrou que eventos exógenos afetam bruscamente a bolsa de valores. Desta forma, outra possível abordagem é a construção de modelos que prevejam os momentos de não se construir carteiras de investimento em momentos de irracionalidade do mercado.

Referências

AALST, W. van der. *Process Mining: Data Science in Action*. 2nd. ed. [S.l.]: Springer, 2016. Citado na página 40.

AGGARWAL, C. *Neural Networks and Deep Learning A Textbook*. [S.l.]: Springer International Publishing AG, 2018. Citado 2 vezes nas páginas 27 e 28.

B3. *Ações*. 2020. <http://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/acoes.htm>. Citado na página 18.

B3. *Histórico*. 2020. <https://ri.b3.com.br/pt-br/b3/historico/>. Citado na página 18.

B3. *Histórico pessoas físicas*. 2020. <http://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/consultas/mercado-a-vista/historico-pessoas-fisicas/>. Citado na página 18.

B3. *Ibovespa B3*. 2020. <http://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/ibovespa.htm/>. Citado na página 20.

B3. *Institucional*. 2020. <http://www.b3.com.br/pt_br/b3/institucional/quem-somos/>. Citado na página 14.

B3. *Cotações*. 2021. <http://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/cotacoes/>. Citado 2 vezes nas páginas 8 e 20.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. (Information Science and Statistics). Citado 4 vezes nas páginas 8, 14, 26 e 27.

BOCHMAN, A. *The Kelly Criterion: You Don't Know the Half of It*. [S.l.]: CFA Institute, 2018. <https://blogs.cfainstitute.org/investor/2018/06/14/the-kelly-criterion-you-dont-know-the-half-of-it/>. Citado 2 vezes nas páginas 24 e 25.

BRAND, T. F. *Artificial Intelligence and The Banking Industry's $1 Trillion Opportunity*. 2018. <https://thefinancialbrand.com/72653/artificial-intelligence-trends-banking-industry>. Citado na página 14.

BROCKMAN, G. et al. *OpenAI Gym*. 2016. Cite arxiv:1606.01540. Disponível em: <http://arxiv.org/abs/1606.01540>. Citado 5 vezes nas páginas 8, 16, 31, 37 e 49.

BROWNLEE, J. *Train-Test Split for Evaluating Machine Learning Algorithms*. [S.l.]: Machine Learning Mastery, 2020. <https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/>. Citado na página 45.

CARTA, A.; CONVERSANO, C. Practical implementation of the kelly criterion: Optimal growth rate, number of trades, and rebalancing frequency for equity portfolios. *Frontiers in Applied Mathematics and Statistics*, 2020. Citado na página 25.

CONEGUNDES, L.; PEREIRA, A. Beating the stock market with a deep reinforcement learning day trading system. *International Joint Conference on Neural Networks*, 2020. Citado 2 vezes nas páginas 15 e 49.

CVM. *O que é uma ação?* [S.l.]: Portal do Investidor - Gov BR, s.d. <https://www.investidor.gov.br/menu/Menu_Investidor/valores_mobiliarios/Acoes/o_que_e_uma_acao.html>. Citado na página 18.

CVM. *Quanto vale uma ação?* [S.l.]: Portal do Investidor - Gov BR, s.d. <https://www.investidor.gov.br/menu/Menu_Investidor/valores_mobiliarios/Acoes/quanto_vale_uma_acao.html>. Citado na página 14.

DENG, Y. et al. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 2017. Citado na página 35.

EILERS, D. et al. Intelligent trading of seasonal effects: A decision support algorithm based on reinforcement learning. *Decision Support Systems*, 2014. Citado na página 35.

ELDER, A. *Come into my trading room*. [S.l.]: John Wiley Sons, Inc., 2002. Citado na página 19.

ELTON, E. et al. *Modern Portfolio Theory and Investment Analysis*. 9th. ed. [S.l.: s.n.], 2014. Citado 5 vezes nas páginas 8, 14, 19, 20 e 21.

FISCHER, T. Reinforcement learning in financial markets - a survey. *FAU Discussion Papers in Economics*, 2018. Citado 3 vezes nas páginas 15, 34 e 35.

GUTHRIE, D. *Yahooquery Python's Package*. [S.l.]: GitHub, 2020. <https://github.com/dpguthrie/yahooquery>. Citado 3 vezes nas páginas 6, 16 e 37.

H., Y. et al. Deep reinforcement learning for automated stock trading: An ensemble strategy. 2020. Citado 3 vezes nas páginas 15, 49 e 62.

HAARNOJA, T. et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *CoRR*, abs/1801.01290, 2018. Disponível em: <http://arxiv.org/abs/1801.01290>. Citado 2 vezes nas páginas 31 e 62.

HARRIS, C. R. et al. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>. Citado na página 37.

HILL, A. et al. *Stable Baselines*. [S.l.]: GitHub, 2018. <https://github.com/hill-a/stable-baselines>. Citado 2 vezes nas páginas 16 e 37.

HUANG, J.-J.; TZENG, G.-H.; ONG, C.-S. Marketing segmentation using support vector clustering. *Expert Systems with Applications*, 2007. Citado na página 26.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. Citado na página 37.

JEWELL, C. *Artificial intelligence: the new electricity*. [S.l.]: World Intellectual Property Organization WIPO - Magazine, 2019. <https://www.wipo.int/wipo_magazine/en/2019/03/article_0001.html>. Citado 3 vezes nas páginas 6, 7 e 26.

KELLY, J. A new interpretation of information rate. 1956. Citado na página 14.

LILLICRAP, T. P. et al. *Continuous control with deep reinforcement learning*. 2019. Citado 4 vezes nas páginas 8, 15, 32 e 33.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008. Citado na página 26.

MACLEAN, L.; ED. *The Kelly Capital Growth Investment Criterion*. [S.l.]: Springer, 2006. (Information Science and Statistics). Citado 6 vezes nas páginas 6, 7, 15, 24, 25 e 63.

MAIONE, C.; NELSON, D.; BARBOSA, R. Research on social data by means of cluster. *Applied Computing and Informatics*, 2019. Citado na página 26.

MANGRAM, M. A simplified perspective of the markowitz portfolio theory. *Global Journal of Business Research*, 2013. Citado 4 vezes nas páginas 14, 15, 23 e 54.

MARKOWITZ, H. *Portfolio Selection*. [S.l.: s.n.], 1952. Citado 4 vezes nas páginas 6, 14, 21 e 23.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 56 – 61. Citado na página 37.

MNIH, V. et al. Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783, 2016. Disponível em: <http://arxiv.org/abs/1602.01783>. Citado na página 62.

NEKRASOV, V. Kelly criterion for multivariate portfolios: A model-free approach. 2014. Citado na página 25.

OLANIYI, A.; ADEWOLE, K.; JIMOH, R. Stock trend prediction using regression analysis - a data mining approach. *ARPN Journal of Systems and Software*, 2011. Citado na página 27.

OLIVEIRA, F.; NOBRE, C.; ZáRATE, L. Applying artificial neural networks to prediction of stock price and improvement of the directional prediction index – case study of petr4, petrobras, brazil. *Expert System with Applications*, 2013. Citado 2 vezes nas páginas 15 e 34.

OPENAI. *Proximal Policy Optimization*. [S.l.]: OpenAI Spinning Up. <https://spinningup.openai.com/en/latest/algorithms/ppo.html>. Citado 3 vezes nas páginas 8, 33 e 34.

OTTERLO, M. van; WIERING, M. Reinforcement learning and markov decision processes. 2012. Citado na página 30.

PALANISAMY, S. *Association Rule Based Classification*. Dissertação (Mestrado) — Worcester Polytechnic Institute, 2006. Citado na página 45.

PLOTLY. *Markowitz Portfolio Optimization in Python/v3*. [S.l.]: Plotly, 2020. <https://plotly.com/python/v3/ipython-notebooks/markowitz-portfolio-optimization/>. Accessed: 2021-02-14. Citado 2 vezes nas páginas 8 e 22.

ROSSUM, G. V.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697. Citado 4 vezes nas páginas 6, 7, 16 e 37.

SATCHELL, S.; SCOWCROFT, A. *Advances in Portfolio Construction and Implementation*. [S.l.]: Elsevier, 2003. Citado na página 21.

SCHULMAN, J. et al. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. Disponível em: <http://arxiv.org/abs/1707.06347>. Citado na página 32.

SHARMA, S.; ATHAIYA, A. Activation functions in neural networks. *International Journal of Engineering Applied Sciences and Technology*, 2020. Citado na página 28.

SHARPE, W. F. The sharpe ratio. *The Journal of Portfolio Management*, Institutional Investor Journals Umbrella, v. 21, n. 1, p. 49–58, 1994. ISSN 0095-4918. Disponível em: <https://jpm.pm-research.com/content/21/1/49>. Citado na página 22.

SILVER, D. *Markov Decision Processes - Lecture 3 - UCL Course on RL*. [S.l.]: David Silver's Blog, 2015. <https://www.davidsilver.uk/teaching/>. Citado 2 vezes nas páginas 29 e 30.

SILVER, D. *Markov Decision Processes - Lecture 5 - UCL Course on RL*. [S.l.]: David Silver's Blog, 2015. <https://www.davidsilver.uk/teaching/>. Citado na página 31.

SUTTON, R.; BARTO, A. *Reinforcement learning: an introduction*. 2nd. ed. [S.l.]: The MIT Press, 2020. (Adaptive Computation and Machine Learning). Citado 10 vezes nas páginas 6, 7, 8, 15, 28, 29, 30, 31, 32 e 33.

Apêndices

APÊNDICE A – Simulações de projeção da valorização da carteira para o período de teste

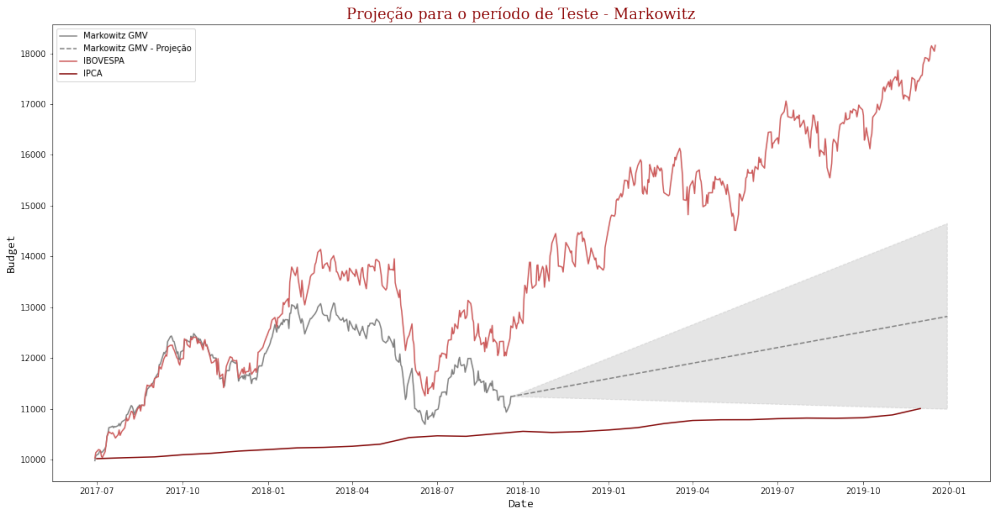


Figura 24 – Projeção da carteira de Markowits para o período de Testes

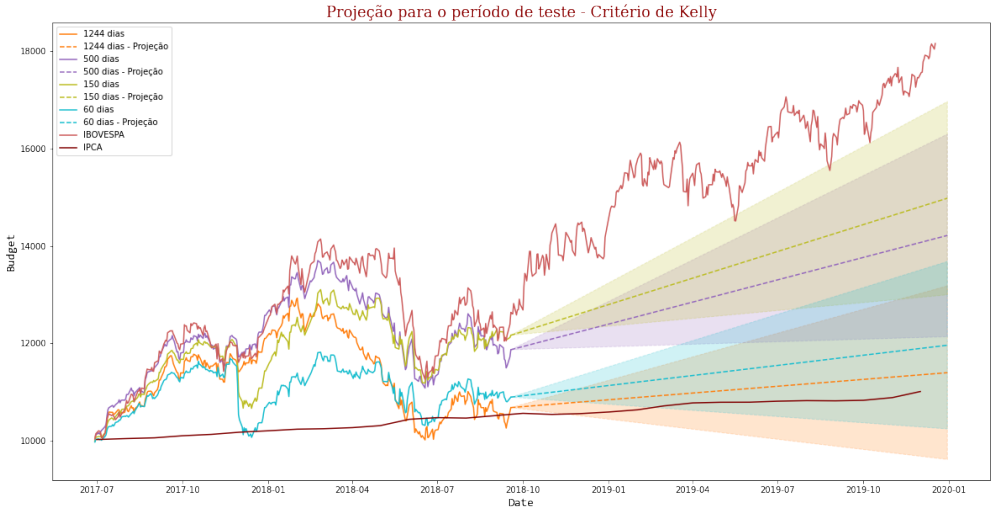


Figura 25 – Projeção das carteiras de Kelly para o período de Testes

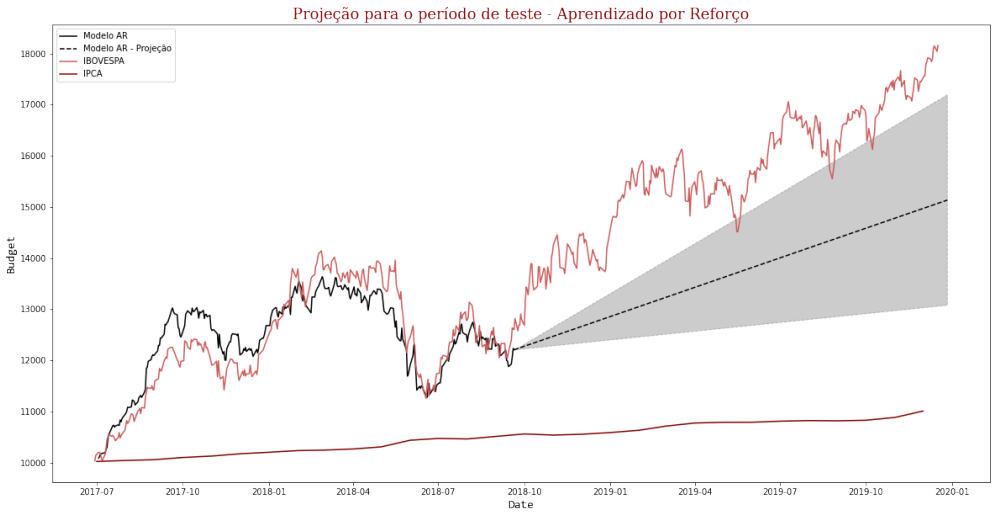


Figura 26 – Projeção da carteira por Aprendizado por Reforço para o período de Testes