MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Colegiado de Engenharia de Produção

Aplicação de aprendizado por reforço para solução do problema dinâmico de gestão de estoque de medicamentos em hospitais com seleção de fornecedores e descontos no frete

Louis Guilherme Marinho de Resende

João Monlevade, MG
2024

Louis Guilherme Marinho de Resende

Aplicação de aprendizado por reforço para solução do problema dinâmico de gestão de estoque de medicamentos em hospitais com seleção de fornecedores e descontos no frete

Trabalho de conclusão de curso apresentado ao curso de Engenharia de Produção do Instituto de Ciências Exatas e Aplicadas da Universidade Federal de Ouro Preto, como parte dos requisitos necessários para a obtenção do título de Bacharel em Engenharia de Produção.

Orientador: Prof. Dr. Thiago Augusto de Oliveira Silva

João Monlevade, MG

2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

R433a Resende, Louis Guilherme Marinho de.
Aplicação de aprendizado por reforço para solução do problema dinâmico de gestão de estoque de medicamentos em hospitais com seleção de fornecedores e descontos no frete. [manuscrito] / Louis Guilherme Marinho de Resende. - 2024.
40 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Thiago Augusto de Oliveira Silva.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Engenharia de Produção .

1. Aprendizado do computador. 2. Controle de estoque - Medicamentos. 3. Hospitais - Custo operacional. 4. Logística - Saúde. 5. Otimização matemática. 6. Simulação (Computadores). I. Silva, Thiago Augusto de Oliveira. II. Universidade Federal de Ouro Preto. III. Título.

CDU 658.5:004.85

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE ENGENHARIA DE PRODUCAO - ICEA

FOLHA DE APROVAÇÃO

Louis Guilherme Marinho de Resende

Aplicação de aprendizado por reforço para solução do problema dinâmico de gestão de estoque de medicamentos em hospitais com seleção de fornecedores e descontos no frete

Monografia apresentada ao Curso de Graduação em Engenharia de Produção da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Engenheiro de Produção.

Aprovada em 11 de outubro de 2024

Membros da banca

Dr. Thiago Augusto de Oliveira Silva - Orientador(a) - Universidade Federal de Ouro Preto
Me. Matheus Correia Teixeira - Charter Communications, USA
Dra. Mônica do Amaral - Universidade Federal de Ouro Preto
Dr. Paganini Barcellos de Oliveira - Universidade Federal de Ouro Preto

Thiago Augusto de Oliveira Silva, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 17/10/2024

Documento assinado eletronicamente por **Thiago Augusto de Oliveira Silva**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 17/10/2024, às 16:10, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0796780** e o código CRC **B0DCBD7A**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0817 - www.ufop.br

Agradecimentos

Gostaria, em primeiro lugar, de expressar a minha profunda gratidão à minha família, que me apoiou ao longo de todos os anos de estudos. Seu amor, incentivo e suporte constante foram pilares essenciais para o meu crescimento pessoal e acadêmico.

Também gostaria de agradecer ao meu tutor, Thiago Augusto, pela orientação ao longo de toda a minha caminhada acadêmica. Seu constante apoio, paciência e seus conselhos sábios foram determinantes para a concretização deste ciclo.

Agradeço ainda aos demais professores e técnicos da UFOP, cujos ensinamentos e orientações ao longo da minha formação foram fundamentais.

Finalmente, gostaria de expressar minha gratidão aos amigos que fiz durante essa jornada, pelo companheirismo e apoio nas diferentes etapas do curso.

Resumo

Este trabalho foca na otimização da gestão de estoques de medicamentos em ambientes hospitalares, reconhecendo a importância vital dos serviços de saúde e as complexas demandas logísticas associadas. O objetivo é desenvolver um modelo que melhore a eficiência na gestão de estoques, levando em consideração a variabilidade e incertezas típicas desse ambiente. Foi proposto um modelo de otimização que combina técnicas de aprendizado por reforço, especificamente o algoritmo *Proximal Policy Optimization* (PPO), com um problema clássico de controle de estoque. O modelo foi avaliado em cenários simulados com diferentes configurações de parâmetros que influenciam os custos operacionais. Os resultados demonstraram que o uso de aprendizado por reforço superou a política determinística — uma abordagem que toma decisões baseadas apenas no estado atual, sem considerar impactos futuros — ao reduzir significativamente os custos totais e melhorar o controle de estoque em condições de incerteza.

Palavras-chave: Dimensionamento de Lotes. Otimização. Aprendizado por Reforço. Simulação. Logística em Saúde.

Abstract

This work focuses on the optimization of medication inventory management in hospital environments, recognizing the critical importance of healthcare services and the complex logistical demands associated with them. The objective is to develop a model that improves inventory management efficiency, considering the variability and uncertainties typical of such environments. An optimization model was proposed, combining reinforcement learning techniques, specifically the Proximal Policy Optimization (PPO) algorithm, with a classical inventory control problem. The model was evaluated in simulated scenarios with different parameter configurations that influence operational costs. The results demonstrated that the use of reinforcement learning outperformed the deterministic policy — an approach that makes decisions based solely on the current state, without considering future impacts — by significantly reducing total costs and improving inventory control under uncertain conditions.

Keywords: *Lot-Sizing*. *Optimization*. *Renforcement Learning*. *Simulation*. *Healthcare logistics*.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A relevância dos serviços de saúde na sociedade contemporânea é incontestável. O desafio de fornecer serviços de maior qualidade a custos reduzidos tem se destacado como uma das principais preocupações nas áreas de engenharia e gestão. De acordo com Rais e Viana (2011), muitos países têm adotado o planejamento em sistemas de saúde como uma estratégia para oferecer serviços adequados. Os hospitais, por sua vez, enfrentam a necessidade de reduzir custos e, ao mesmo tempo, melhorar o atendimento e a satisfação dos pacientes (CARDOEN; DEMEULEMEESTER; BELIËN, 2010; VRIES; HUIJSMAN, 2011).

A demanda por medicamentos está diretamente relacionada ao tratamento de pacientes e, devido às incertezas inerentes a esse processo, apresenta uma característica altamente irregular. Independentemente de seu valor, a logística hospitalar precisa garantir o abastecimento contínuo de todos os pontos de distribuição de medicamentos e materiais médico-hospitalares. A capacidade de atender a essas demandas no momento certo e na quantidade correta está intimamente ligada ao sucesso dos tratamentos e, muitas vezes, à própria vida dos pacientes.

Conforme observado por Rais e Viana (2011), a utilização de técnicas de pesquisa operacional como suporte nas tomadas de decisão em ambientes hospitalares tem crescido significativamente. Os principais desafios incluem a previsão de demanda, a seleção de locais para novas unidades hospitalares e a gestão de emergências, visando um atendimento mais eficiente. A implementação de metodologias como a *Lean Healthcare* tem sido amplamente discutida por diversos autores (ver, por exemplo, Selau et al. (2009) e Bertani (2012)), uma vez que se mostra eficaz na eliminação de desperdícios e na melhoria dos processos. Para que essa metodologia seja implementada com sucesso, é fundamental uma gestão eficiente dos materiais dentro das instituições de saúde.

A gestão de medicamentos, frequentemente realizada pelas farmácias hospitalares, desempenha um papel estratégico na redução de custos e na garantia da qualidade do atendimento ao paciente, como destacado por Gonçalves, NOVAES e Simonetti (2006). Além dos custos associados à falta de medicamentos, os custos de armazenamento e manutenção podem ser consideráveis, devido ao alto valor dos itens, ao risco de obsolescência e às exigências de condições especiais de estocagem. Kritchanchai e Meesamut (2015) observam que, em muitos países em desenvolvimento, os hospitais mantêm estoques excessivos para assegurar níveis elevados de serviço, resultando em despesas desnecessárias.

Diante desse cenário desafiador, esta pesquisa propõe a exploração de uma abordagem que seja capaz de capturar a dinâmica e a aleatoriedade inerentes ao sistema de gestão de medicamentos em ambientes hospitalares. A complexidade da demanda, frequentemente imprevisível, e a variabilidade nos padrões de consumo dos medicamentos reforçam a necessidade de um método que vá além das abordagens tradicionais, sensível às características estocásticas desse tipo de problema.

Embora a gestão de estoque em ambientes hospitalares estabeleça o contexto do problema, o foco deste trabalho reside na estratégia de solução adotada. A pesquisa explora o uso de técnicas avançadas de Aprendizado por Reforço, do inglês *Reinforcement Learning* (RL), com ênfase no PPO, como uma abordagem promissora para enfrentar a variabilidade e incerteza típicas desse cenário. A escolha por essa técnica visa proporcionar maior flexibilidade e adaptabilidade na tomada de decisão em tempo real, permitindo que o sistema se ajuste dinamicamente às flutuações da demanda e aos custos operacionais. Com isso, o trabalho não apenas aborda a gestão de estoques, mas contribui para a inovação nas metodologias aplicadas à otimização de processos hospitalares, fornecendo uma solução capaz de melhorar a eficiência e reduzir custos no longo prazo.

1.1 Objetivo geral

Propor um método de solução baseado em RL que seja capaz de lidar com as incertezas e a dinamicidade da gestão de estoques no contexto hospitalar.

1.1.1 Objetivos específicos

- Construir um modelo que incorpore de maneira precisa as características específicas do problema, considerando a capacidade e os custos operacionais, proporcionando uma base sólida para análises e soluções subsequentes;
- Desenvolver políticas de solução utilizando a abordagem de RL, explorando estratégias que permitam ao sistema aprender e adaptar suas ações ao ambiente dinâmico;
- Realizar uma comparação do desempenho das políticas de solução desenvolvidas, utilizando um modelo de simulação para avaliar a eficácia e a robustez dessas políticas em diferentes cenários, proporcionando *insights* para as tomadas de decisão práticas.

1.2 Organização do Trabalho

Este trabalho está organizado em capítulos que abordam aspectos essenciais da pesquisa: o Capítulo 1 apresenta o contexto, a relevância do problema e define os objetivos da pesquisa; o Capítulo 2 revisa a literatura sobre modelos de estoque com ênfase em modelos dinâmicos, identificando lacunas e oportunidades; no Capítulo 3, é descrita a metodologia utilizada, desde a revisão de literatura até a aplicação de ferramentas computacionais; o Capítulo 4 detalha a modelagem do problema, incluindo a implementação do agente de decisão; o Capítulo 5 apresenta os resultados obtidos nas simulações e as comparações entre as políticas propostas; finalmente, o Capítulo 6 traz as considerações finais, destacando as contribuições, limitações e sugestões para pesquisas futuras.

2 Revisão da Literatura

2.1 Problema de dimensionamento de lote multiproduto com seleção de fornecedores e desconto no frete

O artigo de Basnet e Leung (2005) proporciona uma contribuição para a literatura ao apresentar um modelo para o desafiador problema de dimensionamento de lote de estoque multiproduto e multi-período com seleção de fornecedores. Originalmente, esse problema foi abordado com um algoritmo de enumeração exaustiva e um heurístico. No entanto, os autores introduziram uma abordagem, utilizando técnicas de otimização baseadas em redução e propondo uma nova desigualdade válida que melhorou os resultados numéricos. Assim, o modelo superou as soluções previamente publicadas em diversas instâncias de teste, reafirmando a importância de metodologias avançadas para esse tipo de problema.

Em termos de formulação, Basnet e Leung (2005) propuseram um modelo de Programação Linear Inteira Mista (*Mixed-Integer Linear Programming*) para o problema de dimensionamento de lote de estoque multiperíodo e multiproduto com seleção de fornecedores. A seguir, a Tabela 1 apresenta a notação e a formulação matemática deste artigo:

Tabela 1 – Caracterização dos parâmetros e das variáveis do problema base.

Símbolo	Descrição	Domínio
Conjuntos		
I	Conjunto de produtos i	$\geq 0 \in \mathbb{Z}$
J	Conjunto de fornecedores j	$\geq 0 \in \mathbb{Z}$
T	Conjunto de períodos t	$\geq 0 \in \mathbb{Z}$
Parâmetros		
D_{it}	Demanda do produto i no período t	≥ 0
P_{ij}	Preço de compra ($) do produto i do fornecedor j	≥ 0
H_i	Custo de manutenção ($) do produto i por período	≥ 0
O_j	Custo de pedido ($) do fornecedor j	≥ 0
Variáveis		
X_{ijt}	Tamanho do lote do produto i encomendado ao fornecedor j no período t	≥ 0
Y_{jt}	1 se um pedido for feito ao fornecedor j no período t, 0 caso contrário	$\{0,1\}$

O problema é então formulado como:

$$\min \sum_{t\in T}\sum_{j\in J}\sum_{i\in I} P_{ij}X_{ijt} + \sum_{j\in J}\sum_{t\in T} O_j Y_{jt} + \sum_{i\in I}\sum_{t\in T} H_i \left(\sum_{k=1}^{t}\sum_{j\in J} X_{ijk} - \sum_{k=1}^{t} D_{ik} \right) \tag{2.1}$$

$$\text{s. a.: } \textstyle\sum_{k=1}^{t}\sum_{j\in J} X_{ijk} - \sum_{k=1}^{t} D_{ik} \geq 0, \quad \forall i,t \tag{2.2}$$

$$\left(\textstyle\sum_{k=1}^{t} D_{ik}\right) Y_{jt} - X_{ijk} \geq 0, \quad \forall i,j,t \tag{2.3}$$

$$Y_{jt} = \{0,1\}, \quad \forall j,t \tag{2.4}$$

$$X_{ijk} \geq 0, \quad \forall i,j,t \tag{2.5}$$

A função objetivo (2.1) é o custo total incorrido pelo comprador e compreende o custo total de compra dos produtos, o custo total de pedido e o custo total de manutenção para manter estoque em cada período. As restrições (2.2) garantem que toda a demanda seja satisfeita no período em que ocorre; essas restrições também garantem evitar a escassez. As restrições (2.3) estabelecem que não é possível fazer um pedido sem cobrar um custo de pedido correspondente. As restrições (2.4) definem as variáveis binárias. As restrições (2.5) impõem condições de não negatividade às restantes variáveis de decisão. O problema descrito por (2.1)–(2.5) é, segundo Basnet e Leung (2005), da classe NP-difícil, o que dificulta a solução de instâncias de grande porte.

Oliveira, Souza e Silva (2019) ampliam o escopo da pesquisa ao aplicar técnicas de otimização ao contexto hospitalar, dada a natureza sensível das atividades e a necessidade de manter elevados níveis de serviço. O trabalho apresenta uma metodologia voltada para encontrar limites inferiores de um modelo matemático de dimensionamento de lotes com seleção de fornecedores em farmácias hospitalares. Para isso, propõe-se a implementação da decomposição de Dantzig-Wolfe com um método de geração de colunas para o modelo. Os resultados obtidos pela metodologia revelam a obtenção de limites inferiores superiores aos da relaxação linear, evidenciando a eficácia da abordagem proposta para melhorar a gestão de estoques.

2.2 Abordagens dinâmicas

Na vasta literatura sobre gestão de estoques, diversos estudos enfrentam o desafio de lidar com o planejamento e controle em ambientes dinâmicos. Nessa área é o trabalho de Zipkin (1986), que oferece uma contribuição ao demonstrar que um dos resultados fundamentais da teoria de estoques se mantém válido sob condições mais amplas do que as até então abordadas. Este resultado caracteriza as distribuições dos níveis de estoque e da posição de estoque no modelo padrão de tempo contínuo com pedidos em atraso, permitindo o cálculo relativamente simples de medidas-chave de desempenho. O autor examina os tempos de ressuprimento fixos quanto aleatórios, além de distribuições estacionárias e limitantes sob diferentes suposições. O cerne desse estudo é a consideração de processos de demanda descritos por diversas classes gerais de processos de contagem composta, juntamente com uma variedade de políticas de pedido. Para o caso com tempos de ressuprimento estocásticos, Zipkin (1986) apresenta a primeira prova explícita do resultado, assumindo um cenário plausível para a geração dos tempos de ressuprimento.

Seguindo essa linha de investigação, outros trabalhos avançaram o estado da arte no tratamento de incertezas no planejamento de produção e controle de estoques. Por exemplo, Feiring e Sastri (1990) foca no planejamento de produção sob demanda normalmente distribuída, utilizando programação estocástica para atender a demanda ao menor custo de produção e estoque possível. Em um contexto similar, Anupindi e Tayur (1998) desenvolvem um modelo de gerenciamento de produção de vários itens, buscando minimizar diversos custos operacionais em um ambiente com taxas de produção finitas e tempos de transição entre produtos.

De maneira complementar, Fransoo, Sridharan e Bertrand (1995) propõem uma abordagem hierárquica para o planejamento e programação em sistemas de produção com múltiplos itens e máquinas únicas, levando em conta a incerteza nas demandas. Nesse sentido, Karmarkar e Yoo (1994) lidam com o problema de sequenciação de múltiplos produtos em uma única máquina, utilizando decomposições Lagrangeanas e métodos de limites inferior e superior para atender a demandas estocásticas de forma eficiente.

Em outra vertente, Martel, Diaby e Boctor (1995) trata da aquisição de múltiplos itens com demandas estocásticas não estacionárias, reduzindo o problema a um programa linear 0-1 para cada período de planejamento, facilitando a sua solução. Já Iida (1999) enfrenta o problema de inventário com horizonte infinito e demandas não estacionárias, utilizando uma abordagem de revisão periódica dinâmica. Adicionalmente, Sobel e Zhang (2001) apresentam um sistema de inventário de revisão periódica que considera tanto demandas determinísticas quanto aleatórias, aplicando técnicas de programação dinâmica.

Esses estudos, embora diversos em suas abordagens, compartilham o objetivo comum de aprimorar a gestão de estoques em ambientes dinâmicos e incertos. Cada um traz importantes contribuições para o avanço das técnicas de modelagem e otimização, proporcionando soluções mais eficientes e realistas para problemas de estoque, que continuam sendo fundamentais em diversos setores produtivos.

2.3 Aprendizado por reforço

O RL é uma área de pesquisa em destaque, especialmente devido aos avanços nos algoritmos de aprendizado profundo. Trata-se de uma técnica de Aprendizado de Máquina que capacita o *software* a tomar decisões visando os melhores resultados, emulando o processo de aprendizado por tentativa e erro utilizado pelos seres humanos. As ações do agente de decisão que alcançam seus objetivos são reforçadas, enquanto aquelas que prejudicam a meta são desconsideradas. Sutton e Barto (1998) foram os percussores da área.

Os algoritmos de RL operam com base em um paradigma de recompensa e punição ao processar dados. Eles aprendem com o *feedback* de cada ação, descobrindo os melhores caminhos de processamento para atingir os resultados finais. Esses algoritmos também são capazes de atrasar a gratificação, reconhecendo que a estratégia ótima pode envolver sacrifícios de curto prazo.

Powell (2021) propõe um *framework* de modelagem de cinco elementos que descreve qualquer problema de decisão sequencial, ao enquadrar os problemas em duas classes amplas: problemas de aprendizado puro e problemas gerais dependentes do estado. Ao otimizar políticas, Powell identifica quatro classes de políticas que englobam qualquer método de tomada de decisão. Essa perspectiva moderna do RL questiona se ele é, de fato, um subconjunto desse *framework* mais amplo. Apesar de intitular seu livro *"Reinforcement Learning and Stochastic Optimization"*, Powell sugere que *"Sequential Decision Analytics"* poderia ter sido uma denominação mais apropriada, abrangendo algoritmos de gradientes de políticas, modelagem determinística de otimização e políticas baseadas em aproximações de antecipação. O livro oferece uma visão abrangente de modelos determinísticos e estocásticos, proporcionando *insights* para aplicação em diversos contextos, como a gestão de estoque de medicamentos em hospitais com seleção de fornecedores, ao explorar diferentes classes de políticas e suas hibridizações.

A Figura 1 apresenta o ciclo do RL em relação a uma política utilizada na modelagem de um problema e ao ambiente de simulação. Esse ciclo ilustra como o agente interage com o ambiente, observando o estado atual, selecionando uma ação com base na política definida, recebendo uma recompensa ou penalidade do ambiente e, em seguida, atualizando sua política com base no *feedback* recebido. Esse processo iterativo de interação contínua com o ambiente é fundamental para que o agente aprenda a tomar decisões que levem aos resultados desejados ao longo do tempo. No contexto específico do problema de gestão de estoque com seleção de fornecedores, esse ciclo de aprendizado pode ser aplicado para determinar as melhores políticas, considerando fatores como demanda variável, prazos de entrega dos fornecedores, custos de armazenamento e restrições de orçamento. O RL oferece uma abordagem dinâmica e adaptativa para lidar com as complexidades desse ambiente, permitindo que o sistema de gestão de estoque tome decisões eficazes e eficientes em tempo real.

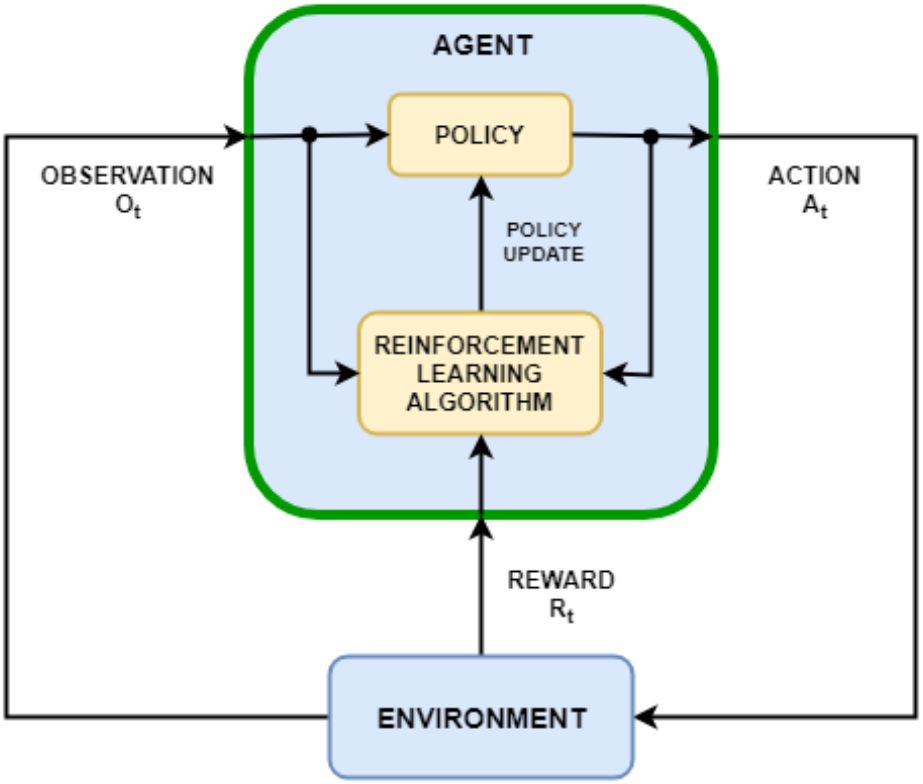

Figura 1 – *Framework* de interação Agente-Ambiente
. (SIRASKAR, 2021)

2.3.1 *Proximal Policy Optimization*

O PPO é um algoritmo de aprendizado por reforço introduzido por Schulman et al. (2017), amplamente reconhecido por sua eficácia e simplicidade de implementação. Como um algoritmo do tipo ator-crítico, o PPO conta com dois componentes principais: o ator, responsável por selecionar as ações, e o crítico, encarregado de avaliar as ações tomadas pelo ator com base em uma função de valor. Diferentemente de outros métodos de gradiente de política, o PPO busca garantir a estabilidade do treinamento ao controlar o tamanho das atualizações da política em cada iteração, de forma a não comprometer o desempenho ou a convergência adequada do algoritmo. Para isso, utiliza uma abordagem conservadora que mede a variação da política em relação à iteração anterior, aplicando um rácio entre as probabilidades de ação das políticas atual e anterior. Esse cálculo é então limitado a um intervalo $(1-\varepsilon, 1+\varepsilon)$, impedindo que a política atual se afaste excessivamente da anterior. Essa característica é conhecida como política próxima (*proximal policy*).

A Figura 2 apresenta a relação entre a ação tomada pela política $\pi_\theta(a_t|s_t,\theta)$ e o valor da função objetivo $L^{CLIP}(\theta)$ no algoritmo de PPO. O algoritmo utiliza duas redes neurais: a rede π_θ, que atua como o **ator**, responsável por selecionar a ação a_t dado o estado s_t, e a rede V_ϕ, que funciona como o **crítico**, estimando o valor esperado do retorno futuro a partir do estado atual s_t.

A função objetivo $L^{CLIP}(\theta)$ ajusta as atualizações da política com base na razão de probabilidade $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$, que compara a política atual com a política anterior, limitando as mudanças excessivas através de um intervalo de *clipping* (*clipping range*) controlado por ε.

Quando a vantagem A_t, que mede o ganho esperado da ação a_t sobre a média das ações possíveis no estado s_t, é positiva (parte esquerda da figura), isso indica que a ação a_t foi melhor do que o esperado. Nesse caso, o objetivo é aumentar a probabilidade de selecionar essa ação novamente, ou seja, $r_t(\theta) > 1$. No entanto, para evitar que a política mude de forma drástica, o PPO aplica o *clipping*, limitando a razão de probabilidade a $1+\varepsilon$, resultando em uma atualização máxima de $(1+\varepsilon)A_t$.

De forma análoga, quando A_t é negativo (parte direita da figura 2), significa que a ação a_t foi pior do que o esperado, e a política deve diminuir a probabilidade de tomar essa ação novamente. Novamente, o *clipping* é aplicado, impedindo que a política mude muito drasticamente, e a atualização é limitada a $(1-\varepsilon)A_t$.

Dessa forma, o algoritmo PPO equilibra a exploração de novas ações com a limitação de mudanças excessivas na política, utilizando o *clipping* para manter a estabilidade do treinamento.

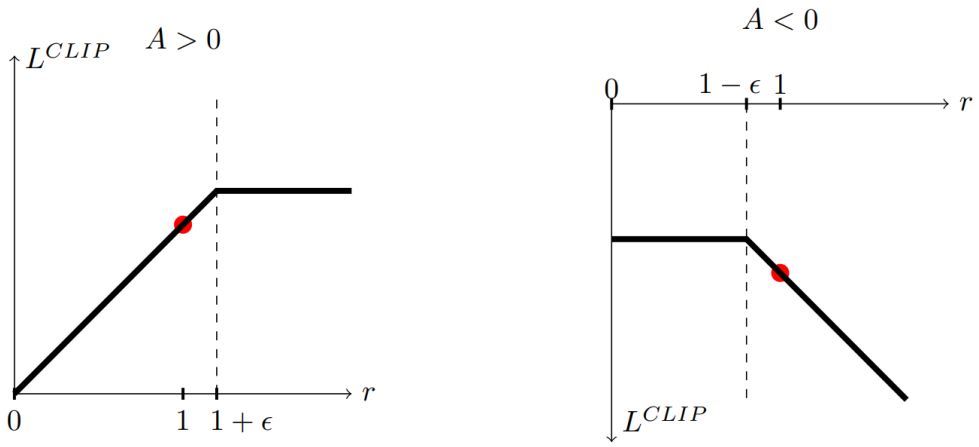

Figura 2 – Gráficos da Função L CLIP. (SCHULMAN et al., 2017)

O algoritmo PPO, mostrado no Algoritmo 1, descreve o processo de atualização da política $\pi_\theta(a|s)$ e da função de valor $V_\phi(s)$ por meio de aprendizado iterativo.

Algoritmo 1: Pseudo-algoritmo do *Proximal Policy Optimization*

Dados: Política $\pi_\theta(a|s)$ e função de valor $V_\phi(s)$

Resultado: Política atualizada $\pi_{\theta'}(a|s)$ e função de valor atualizada $V_{\phi'}(s)$

1 Inicializar parâmetros θ e ϕ;
2 **para** cada época de treinamento **faça**
3 Coletar um conjunto de trajetórias $\{\tau_i\}$;
4 **para** cada trajetória $\tau_i = \{(s_t, a_t, r_t)\}$ **faça**
5 Calcular retornos R_t e vantagem A_t;
6 **fim**
7 Otimizar função objetivo $L(\theta)$ de $\pi_\theta(a|s)$ usando gradiente ascendente;
8 Otimizar função de valor $L_V(\phi)$ de V_ϕ usando gradiente descendente;
9 Atualizar parâmetros θ e ϕ;
10 **fim**

A implementação começa com a inicialização dos parâmetros θ e ϕ (Linha 3), que representam, respectivamente, os pesos da rede neural do ator e do crítico. A política π_θ escolhe ações a com base no estado s, enquanto a função de valor V_ϕ estima o valor esperado do retorno futuro a partir de s.

Em cada época de treinamento (Linha 4), o agente coleta um conjunto de trajetórias $\{\tau_i\}$, onde cada trajetória τ_i é uma sequência de estados s_t, ações a_t, e recompensas r_t coletadas ao longo de vários episódios (Linha 5). Essas trajetórias refletem as interações do agente com o ambiente, representando os dados de experiência utilizados para o aprendizado.

Para cada trajetória τ_i, são calculados os retornos R_t e as vantagens A_t (Linha 6). O retorno R_t representa a soma das recompensas futuras descontadas a partir do tempo t, enquanto a vantagem A_t mede o quanto a ação a_t foi melhor ou pior que o esperado, baseado no valor estimado pelo crítico.

Uma vez calculadas as vantagens e os retornos, o algoritmo passa a otimizar a política π_θ e a função de valor V_ϕ. A função objetivo $L(\theta)$ de π_θ é otimizada usando gradiente ascendente (Linha 7), buscando maximizar as recompensas esperadas. Em seguida, a função de valor $L_V(\phi)$ é otimizada por meio de gradiente descendente (Linha 8), minimizando o erro entre os valores estimados pelo crítico e os retornos reais observados.

Por fim, os parâmetros θ e ϕ são atualizados (Linha 9), completando uma iteração do algoritmo. Esse processo é repetido ao longo das épocas de treinamento, ajustando gradualmente a política e a função de valor até a convergência.

Alguns estudos fizeram uso do RL e PPO para as técnicas de solução, o trabalho apresentado por Cai et al. (2019) aborda a busca por soluções mais eficientes em problemas de otimização combinatória, como o problema NP-difícl de empacotamento, utilizando algoritmos heurísticos e aprendizado por reforço. Tradicionalmente, algoritmos como *Simulated Annealing*, Concorde e METIS são empregados para encontrar soluções nesses problemas, mas enfrentam limitações devido à necessidade de uma alta complexidade amostral para obter resultados satisfatórios a partir de uma partida fria. O novo *framework* proposto, denominado RLHO, visa gerar soluções iniciais mais eficazes para os algoritmos heurísticos, capacitando-os a melhorar uma solução inicial gerada por RL. O estudo demonstra que o RL pode utilizar o desempenho da heurística como um sinal de aprendizagem para gerar uma melhor inicialização. A aplicação dessa estrutura em algoritmos como o PPO e o *Simulated Annealing* resultou em experimentos que mostram a superação das linhas de base estabelecidas. Especificamente, no problema de empacotamento, o RL aprendeu a auxiliar a heurística a obter um desempenho ainda melhor, permitindo a combinação das melhores características de ambas as abordagens.

O estudo mais recente conduzido por Hezewijk et al. (2023) investiga o problema estocástico de dimensionamento de lote com capacidade para múltiplos itens e demanda estacionária, com o objetivo de minimizar os custos de configuração, armazenagem e *backorder*. A pesquisa avalia a eficiência do algoritmo PPO nesse contexto, formulando-o como um Processo de Decisão de Markov, do inglês *Markov Decision Process* (MDP) que pode ser resolvido idealmente em cenários menores usando Programação Dinâmica. Os resultados mostram que o desempenho do PPO se aproxima da solução ideal em configurações menores. No entanto, em problemas maiores, a solução ótima torna-se difícil de ser encontrada em tempo computacional viável devido à sua complexidade. Nessas circunstâncias, o estudo revela que a solução do PPO supera uma solução de referência, propondo adaptações ao algoritmo padrão para torná-lo mais escalável em problemas maiores. Além disso, o estudo analisa o crescimento linear no tempo de computação do algoritmo e propõe métodos para explicar seus resultados. Como perspectiva futura, são sugeridas direções de pesquisa que visam melhorar tanto a escalabilidade quanto a explicabilidade do algoritmo PPO.

3 Metodologia

Esta pesquisa é classificada como empírica normativa baseada em modelagem quantitativa. A abordagem segue uma natureza aplicada, centrada na geração de conhecimento para solução de problemas específicos. Quanto ao método, é qualificado como quantitativo, sendo que, conforme Rodrigues (2007), esse tipo de método traduz em números as opiniões e informações para serem classificadas e analisadas.

3.1 Etapas da pesquisa

O processo metodológico iniciou-se com uma análise documental, que consistiu em uma revisão sucinta das publicações relacionadas ao tema. Esse levantamento foi essencial para garantir a relevância do estudo, evitando redundâncias e fornecendo uma visão abrangente das abordagens existentes.

Na fase de implementação e análise, foram empregadas ferramentas e *softwares* especializados para o desenvolvimento, simulação e otimização do modelo. A linguagem de programação escolhida foi o *Python*, juntamente com bibliotecas específicas que suportam algoritmos de otimização, simulação e aprendizado por reforço.

3.1.1 Ferramentas e Pacotes Utilizados

- **Google Colab**: Utilizado como o ambiente principal para o desenvolvimento e execução dos experimentos, o *Google Colab* oferece recursos de processamento remoto, suporte a *GPU* (Unidade de Processamento Gráfico, do inglês *Graphics Processing Unit*) para acelerar operações em redes neurais e outros cálculos paralelos de alto desempenho, além de fácil integração com bibliotecas. Isso viabiliza a execução eficiente das simulações e o treinamento dos modelos em larga escala.

- **Google Drive**: O *Google Drive* foi montado no ambiente para armazenar os resultados das simulações e os modelos treinados, garantindo a persistência dos dados e permitindo o carregamento de políticas previamente treinadas para novos testes e análises.

- **Gurobi**: A biblioteca *Gurobipy*, a interface Python do solver de otimização *Gurobi*, foi utilizada para resolver o problema de otimização relacionado às decisões de reabastecimento. Integrada ao ambiente de simulação, ela assegura que as decisões tomadas sejam otimizadas conforme os parâmetros definidos no modelo. (Gurobi Optimization, LLC, 2023).

- **Stable Baselines3**: Para a implementação dos algoritmos de aprendizado por reforço, foi utilizada a biblioteca *Stable Baselines3*, que fornece uma coleção robusta de algoritmos para problemas de decisão sequencial, com destaque para o algoritmo PPO. (RAFFIN et al., 2021).

- **Torch**: A biblioteca *PyTorch* foi empregada para a implementação das redes neurais no treinamento do algoritmo PPO, garantindo suporte eficiente para operações de alto desempenho, especialmente ao utilizar GPUs. (PyTorch Developers, 2023).

- **Gymnasium**: A simulação do ambiente foi construída com a biblioteca *Gymnasium*, amplamente utilizada para o desenvolvimento de ambientes de aprendizado por reforço, facilitando a integração com os algoritmos de treinamento. (Farama Foundation, 2023).

- **Numpy**: A biblioteca *Numpy* foi essencial para a manipulação de arrays e operações numéricas de alto desempenho, oferecendo uma vasta gama de funções para lidar com os dados gerados pelas simulações. (HARRIS et al., 2020).

- **Pandas**: Utilizada para manipulação e análise de dados, a biblioteca *Pandas* possibilitou a criação e gerenciamento de *dataframes*, facilitando a organização dos dados e a criação de *datasets* durante as simulações. (TEAM, 2023).

- **Scipy**: A biblioteca *Scipy* foi utilizada para realizar cálculos estatísticos durante o processamento dos resultados das simulações. (VIRTANEN et al., 2020).

- **Matplotlib**: A biblioteca *Matplotlib* foi usada para visualização dos dados gerados pelas simulações. Ela permitiu a criação de gráficos como *boxplots* e gráficos de barras. (HUNTER, 2007).

- **Seaborn**: Complementando o *Matplotlib*, a biblioteca *Seaborn* foi utilizada para criar visualizações mais sofisticadas, incluindo gráficos de densidade e outras formas avançadas de visualização. (WASKOM; TEAM, 2023).

3.1.2 *Framework* e Testes

O ambiente de simulação foi configurado para testar diferentes políticas de reabastecimento e avaliar seu impacto sobre métricas como custos de pedido, compra, manutenção de estoque, penalidades por falta e excesso de estoque.

- **Ambiente de Teste**: O ambiente de simulação foi submetido a uma série de testes para garantir sua consistência e robustez, assegurando que ele estivesse corretamente configurado para interagir com os algoritmos de aprendizado por reforço.

- **Políticas Determinísticas vs Políticas Treinadas**: Nos experimentos, foram comparadas políticas determinísticas (com parâmetros fixos) e políticas aprendidas por meio do algoritmo PPO, com o objetivo de avaliar o impacto de cada abordagem no desempenho do sistema.

A implementação do algoritmo PPO no contexto do aprendizado por reforço foi uma escolha estratégica, ampliando a capacidade do modelo em lidar com a complexidade e a dinamicidade do problema de gestão de estoque. Essa abordagem possibilita a otimização contínua das políticas de decisão, adaptando-se às variações de demanda e custos.

Os testes realizados visam avaliar o desempenho do algoritmo PPO no problema modelado, analisando sua eficiência e capacidade de atender às necessidades do ambiente simulado. Os resultados desses testes são discutidos no capítulo de resultados, fornecendo uma análise detalhada da eficácia da abordagem e seu potencial de aplicação em cenários reais.

4 Modelagem

A gestão de inventários é uma área crucial no planejamento logístico e de operações, sendo o problema de quanto comprar um dos principais desafios enfrentados pelas empresas. Essa decisão está diretamente relacionada ao balanceamento entre os custos envolvidos e a incerteza inerente à demanda futura. Uma modelagem adequada permite otimizar as decisões de reabastecimento, levando em consideração diferentes fatores que influenciam tanto os custos quanto os riscos associados à falta ou ao excesso de estoque.

No contexto da decisão de quanto comprar, o principal desafio reside na previsão da demanda ao longo do tempo. Consideramos a demanda como a incerteza inerente ao sistema para a modelagem do problema, já que ela pode variar por diversas razões, incluindo flutuações sazonais, mudanças nos padrões de consumo e eventos imprevistos no mercado. Essa incerteza exige a adoção de políticas robustas que consigam balancear os custos e riscos associados a variações inesperadas no comportamento da demanda.

Os principais tipos de custos que influenciam a decisão de reabastecimento são: os custos de compra ou produção dos produtos, os custos de manutenção de estoque e os custos associados à falta de estoque, também conhecidos como custos de ruptura. Além disso, existem os custos de pedido, que consideram os gastos operacionais e administrativos para realizar um pedido de reabastecimento. Cada um desses componentes deve ser cuidadosamente ponderado ao formular a função objetivo de um problema de otimização de inventário.

Adicionalmente, o problema de gestão de inventários é modelado como uma decisão sequencial, em que a quantidade a ser comprada é decidida periodicamente ao longo de um horizonte temporal. Em alguns casos, assume-se um horizonte infinito, ou seja, o processo de decisão se repete indefinidamente, o que reflete a natureza contínua das operações de reabastecimento. Nessas situações, o objetivo da modelagem é garantir que as políticas de estoque sejam otimizadas não apenas para o curto prazo, mas também ao longo de um período prolongado, de forma a minimizar os custos acumulados ao longo do tempo.

4.1 Interação Agente-Ambiente

A Figura 3 ilustra o ciclo de interação no contexto da modelagem do ambiente de estoque de medicamentos utilizando políticas de reabastecimento baseadas em aprendizado por reforço. Nessa estrutura, o agente, que define as políticas de reabastecimento, interage com o ambiente de inventário e ajusta suas decisões com base nas recompensas recebidas ao longo do tempo. Esse ciclo reflete o processo de aprendizado contínuo, onde o agente busca otimizar suas ações para minimizar os custos operacionais associados à gestão do estoque, levando em consideração as incertezas na demanda e os custos de reabastecimento.

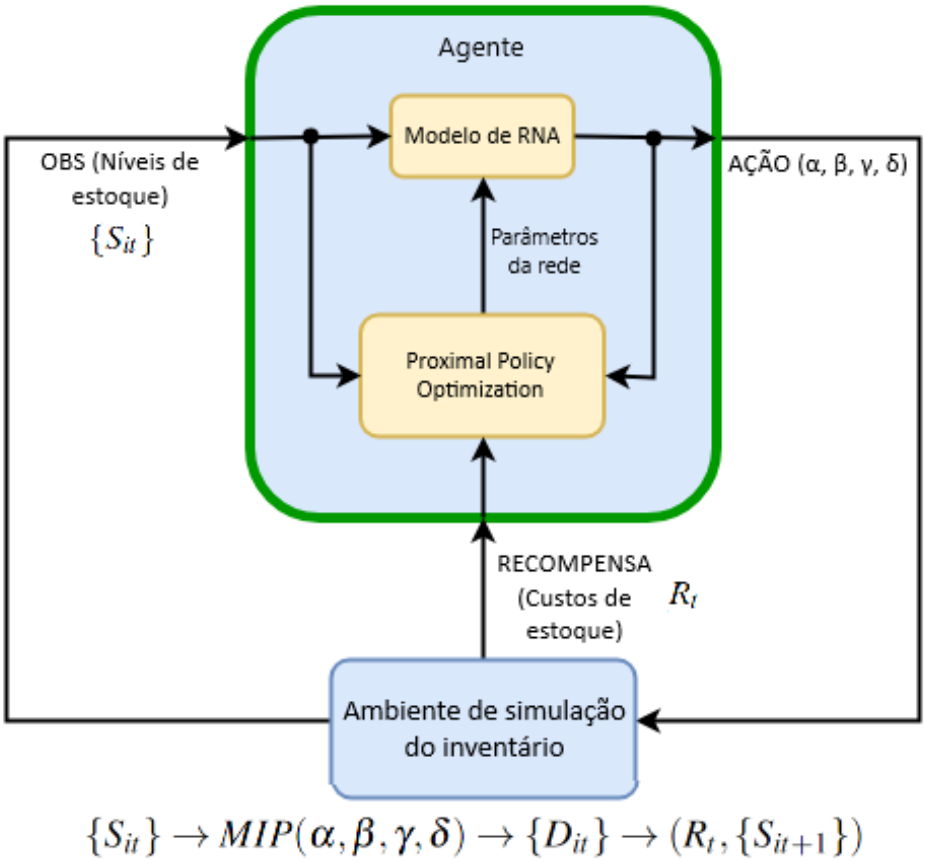

Figura 3 – *Reinforcement Learning Loop adaptado ao ambiente de estoque de medicamentos*. (SIRASKAR, 2021)

Neste contexto, o agente é a entidade responsável por tomar decisões a cada passo do tempo. Ele observa o estado atual do ambiente, que, no caso deste estudo, corresponde ao sistema de gerenciamento de inventário hospitalar. Com base nessas observações, o agente seleciona uma ação (como definir os níveis de reabastecimento de estoque), seguindo uma estratégia determinada por sua política, que é parametrizada e ajustada durante o processo de aprendizado.

Quando uma ação é executada no ambiente, ela gera uma resposta: o ambiente evolui para um novo estado e fornece uma recompensa ao agente. Essa recompensa quantifica o impacto da ação escolhida em termos de custos operacionais, eficiência no gerenciamento do estoque e satisfação da demanda. A função de recompensa guia o agente a preferir ações que levem a um menor custo total ao longo do tempo.

A função **MIP(α, β, γ e δ)**, implementada com o *solver* Gurobi dentro do ambiente, é acionada após a recepção de uma ação composta pelos parâmetros α, β, γ e δ. Esses parâmetros ajustam as restrições e a função objetivo do problema de otimização, permitindo que a função MIP encontre a solução ótima para a quantidade de itens a serem reabastecidos. O resultado dessa otimização gera a recompensa R_t, que reflete os custos operacionais para o período t, e também a nova observação S_{it}, que corresponde aos níveis de estoque ajustados após a ação do agente. Essa interação é fundamental para o aprendizado do agente, pois ele utiliza as recompensas para ajustar sua política ao longo dos episódios simulados.

Esse ciclo se repete de forma contínua ao longo dos episódios de simulação. A cada iteração, o agente ajusta sua política com base nas experiências passadas, utilizando algoritmos como o PPO para maximizar as recompensas esperadas. Dessa forma, o processo de aprendizado por reforço permite que o agente melhore progressivamente seu desempenho ao longo do tempo, aprendendo a equilibrar os diferentes componentes de custo e a otimizar o gerenciamento de estoque.

Assim, o *loop* de aprendizado por reforço resume o processo de tomada de decisão do agente, a interação com o ambiente, a avaliação de ações através de recompensas, e o ajuste iterativo da política, que ocorre até a convergência de uma solução eficiente.

4.1.1 Simulação do Ambiente

O ambiente de simulação foi projetado para replicar o processo de reabastecimento de medicamentos em hospitais, envolvendo múltiplos produtos, fornecedores e períodos. Ele permite testar e avaliar diferentes políticas de controle de estoque, visando minimizar os custos operacionais relacionados ao atendimento de uma demanda gerada aleatoriamente. O objetivo principal é garantir o abastecimento adequado dos medicamentos, evitando tanto o excesso de estoque quanto a falta de produtos essenciais.

O modelo é estruturado como um *Processo de Decisão Markoviano* (MDP), onde as decisões de reabastecimento influenciam a transição entre estados, considerando as incertezas na demanda e no tempo de entrega. As principais características do ambiente incluem:

- **Estados:** Cada estado captura o nível de estoque atual de cada medicamento e o período atual. A transição entre os estados reflete tanto as decisões de reabastecimento quanto a demanda observada ao longo dos períodos.

- **Ações:** As ações correspondem às quantidades de medicamentos a serem encomendadas em cada período. Essas decisões são tomadas pelo agente, que ajusta os parâmetros α, β, γ e δ durante o treinamento. As ações estão relacionadas aos custos e sujeitas a restrições, como o estoque de segurança.

- **Recompensas:** A função de recompensa é definida com base em diferentes custos operacionais, como:

  - **Custo de manutenção:** Relacionado ao armazenamento dos volumes de estoque.
  - **Custo de pedido:** Associado à realização de pedidos aos fornecedores, abrangendo logística e processamento.
  - **Penalidades por falta (fixo e variável):** Aplicadas quando o estoque é insuficiente para atender à demanda, afetando o atendimento hospitalar.
  - **Custo de excesso de estoque:** Refere-se ao desperdício causado por produtos em excesso ou vencidos.

  A recompensa global visa minimizar esses custos, incentivando o agente a equilibrar o atendimento da demanda com o controle de custos operacionais.

- **Demanda:** A demanda é gerada de forma estocástica, com base em distribuições probabilísticas que consideram variações aleatórias no consumo de medicamentos. Isso introduz incerteza no sistema, desafiando o agente a adaptar suas estratégias de reabastecimento.

O ambiente é integrado ao solver de otimização Gurobi, que resolve os problemas de alocação e reabastecimento em cada período, levando em consideração as ações do agente. A cada episódio, uma nova demanda é gerada aleatoriamente, garantindo que o treinamento seja realizado em diferentes cenários de incerteza.

O ambiente segue uma estrutura iterativa, com transições de estado ocorrendo à medida que o agente toma decisões de reabastecimento. A seguir, é apresentado o pseudoalgoritmo 2 que descreve o funcionamento geral do ambiente de simulação.

Algoritmo 2: Pseudoalgoritmo do ambiente de simulação

Dados: Estados iniciais S_0, parâmetros α, β, γ, δ, número de episódios N, demanda histórica d_t, estoque S_t e recompensa R_t

1 **para** cada episódio $t = 1, 2, \ldots, N$ **faça**
2 Observar o estado atual s_t ;
3 Executar MIP(α, β, γ, δ) ;
4 Observar a demanda d_t para o período t ;
5 Atualizar o estado S_{t+1};
6 Retornar a recompensa R_t para o agente;
7 **fim**

O algoritmo começa com a observação dos estados iniciais, que incluem os níveis de estoque e a demanda histórica. A cada iteração, o agente seleciona uma ação de reabastecimento com base na política de controle de estoque. O estado é atualizado após a execução das ações e a observação da demanda, permitindo que o agente refine suas decisões em busca de uma política mais eficiente. Ao longo dos episódios, o ambiente simula diversas situações, permitindo ao agente ajustar seus parâmetros e aprimorar a política de reabastecimento.

4.1.2 Agente

4.1.2.1 Parâmetros da Política

O problema de otimização de inventário é abordado através de uma função objetivo parametrizada por três variáveis principais: α, β e γ, que representam as ponderações atribuídas a três tipos de custo: custo de compra, custo de pedido e custo de manutenção de estoque. Esses parâmetros controlam o comportamento do sistema de gestão de inventário, permitindo ao agente ajustar as políticas de reabastecimento de acordo com as prioridades de minimização de custo.

4.1.2.2 Custo de Compra (α)

O custo de compra refere-se ao valor gasto na aquisição de produtos de fornecedores, representado pelo parâmetro α. Este componente busca minimizar o total gasto com a compra de mercadorias, essencial para reduzir as despesas operacionais da empresa. No entanto, atribuir um peso elevado a α pode induzir o sistema a tomar decisões que favoreçam a redução do custo de compra em detrimento de outros fatores, como a necessidade de reabastecimentos frequentes ou o acúmulo de grandes quantidades de estoque.

$$\alpha \times \sum_{t \in T} \sum_{j \in J} \sum_{i \in I} P_{ij} X_{ijt} \tag{4.1}$$

Na equação acima, P_{ij} é o preço do produto i oferecido pelo fornecedor j, enquanto X_{ijt} denota a quantidade comprada do fornecedor j no período t. O parâmetro α ajusta o peso desse componente na função objetivo, influenciando a prioridade dada ao custo de compra na tomada de decisão.

4.1.2.3 Custo de Pedido (β)

O custo de pedido, ponderado por β, refere-se às despesas fixas associadas à emissão de ordens de compra para os fornecedores. Isso inclui custos administrativos e logísticos, bem como outros encargos relacionados ao processamento de pedidos. Ao otimizar esse componente, o agente tende a emitir um menor número de pedidos, favorecendo a consolidação das compras.

$$\beta \times \sum_{j \in J} \sum_{t \in T} O_j Y_{jt} \tag{4.2}$$

O_j é o custo fixo de emitir um pedido ao fornecedor j, e Y_{jt} é uma variável binária que indica se um pedido foi feito no período t. Aumentar β significa atribuir mais importância à redução da quantidade de pedidos, mesmo que isso possa acarretar maiores custos de compra ou manutenção de estoque.

4.1.2.4 Custo de Manutenção de Estoque (γ)

O custo de manutenção de estoque, ponderado por γ, reflete as despesas associadas ao armazenamento de produtos ao longo do tempo, incluindo custos de espaço, seguro e riscos de depreciação. Um valor elevado de γ incentiva o sistema a manter estoques reduzidos, minimizando os custos de manutenção.

$$\gamma \times \sum_{i \in I} \sum_{t \in T} H_i S_{it} \tag{4.3}$$

Nessa equação, H_i é o custo de manter estoque do produto i, e S_{it} é o nível de estoque do produto i no período t. A ponderação de γ permite que o agente controle o impacto desse custo, buscando um equilíbrio entre a disponibilidade de produtos e a eficiência no gerenciamento de estoque.

4.1.2.5 Restrição de Estoque de Segurança (δ)

Além dos custos, o agente também controla a restrição de estoque de segurança, representada pelo parâmetro δ. Esse parâmetro define o nível mínimo de estoque que deve ser mantido em relação à demanda acumulada, garantindo que a empresa evite rupturas de estoque e seja capaz de atender à demanda futura.

$$S_{it} \geq \delta \times \sum_{k=1}^{t} D_{ik}, \quad \forall i \in I, \forall t \in T \tag{4.4}$$

Nesta expressão, S_{it} representa o estoque do produto i no período t, D_{ik} é a demanda acumulada para o produto i até o período t, e δ define o nível de estoque de segurança em relação à demanda. Um valor maior de δ impõe uma exigência de maior estoque de segurança, o que garante proteção contra incertezas na demanda, mas aumenta os custos de manutenção de estoque.

4.1.3 Modelo determinístico de decisão

O modelo matemático foi ajustado para integração com o ambiente de aprendizado por reforço. Nesse processo, foi implementado um controle de estoque de segurança e adicionados os quatro parâmetros (α, β, γ e δ), que permitem uma maior flexibilidade na definição das políticas de reabastecimento, considerando múltiplos aspectos da operação.

Tabela 2 – Caracterização dos parâmetros e das variáveis do problema base.

Símbolo	Descrição	Domínio
Conjuntos		
I	Conjunto de produtos i	$\geq 0 \in \mathbb{Z}$
J	Conjunto de fornecedores j	$\geq 0 \in \mathbb{Z}$
T	Conjunto de períodos t	$\geq 0 \in \mathbb{Z}$
Parâmetros		
D_{it}	Demanda do produto i no período t	$\geq 0 \in \mathbb{Z}$
P_{ij}	Preço de compra ($) do produto i do fornecedor j	≥ 0
H_i	Custo de manutenção ($) do produto i por período	≥ 0
O_j	Custo de pedido ($) do fornecedor j	≥ 0
α	Proporção de reabastecimento com base na demanda histórica	$\geq 0 \leq 1$
β	Proporção da frequência e quantidade de pedidos de reposição de estoque	$\geq 0 \leq 1$
γ	Proporção dos custos de manutenção de estoque e penalidades por falta	$\geq 0 \leq 1$
δ	Proporção do nível de estoque de segurança	$\geq 0 \leq 1$
Variáveis		
X_{ijt}	Tamanho do lote do produto i encomendado ao fornecedor j no período t	$\geq 0 \in \mathbb{Z}$
Y_{jt}	1 se um pedido for feito ao fornecedor j no período t, 0 caso contrário	$\{0,1\}$
S_i^0	Estoque inicial do produto i	$\geq 0 \in \mathbb{Z}$
S_{it}	Estoque do produto i no período t	$\geq 0 \in \mathbb{Z}$

Então, o problema é modelado da seguinte forma:

$$\min \sum_{t\in T}\sum_{j\in J}\sum_{i\in I} \alpha P_{ij} X_{ijt} + \sum_{j\in J}\sum_{t\in T} \beta O_j Y_{jt} + \sum_{i\in I}\sum_{t\in T} \gamma H_i S_{it} \tag{4.5}$$

$$\text{s. a.:} \quad S_{i1} - S_i^0 - \sum_{j\in J} X_{ij1} + D_{i1} = 0, \qquad \forall i, t = 1 \tag{4.6}$$

$$S_{it} - S_{it-1} - \sum_{j\in J} X_{ijt} + D_{it} = 0, \qquad \forall i, t > 1 \tag{4.7}$$

$$(\sum_{k=1}^{t} D_{ik}) Y_{jt} - X_{ijt} \geq 0, \qquad \forall i, j, t \tag{4.8}$$

$$S_{it} - \delta \sum_{k=1}^{t} D_{ik} \geq 0, \qquad \forall i, t \tag{4.9}$$

$$Y_{jt} \in \{0,1\}, \qquad \forall j, t \tag{4.10}$$

$$X_{ijk} \geq 0, \qquad \forall i, j, t \tag{4.11}$$

Neste contexto, o conjunto T representa os períodos em que as decisões de reabastecimento são tomadas, e não necessariamente todos os períodos disponíveis no horizonte de planejamento. Isso é particularmente importante em ambientes com horizonte infinito, onde o processo de tomada de decisão continua indefinidamente. Neste tipo de configuração, o agente deve considerar os impactos de suas ações não apenas no curto prazo, mas também ao longo de um horizonte de tempo indefinido. Assim, embora as decisões de reabastecimento sejam tomadas em intervalos específicos, o ambiente é projetado para otimizar o desempenho considerando as implicações de longo prazo, balanceando os custos e as recompensas em um ciclo contínuo de aprendizado e adaptação.

A função objetivo apresentada na equação (4.5) representa o custo total do sistema e é composta por três partes principais, cada uma delas associada a um parâmetro da política: α, β e γ. Esses parâmetros permitem ajustar o impacto de diferentes componentes do custo, proporcionando flexibilidade ao tomador de decisão.

O parâmetro α pondera o custo de compra dos produtos ($P_{ij}X_{ijt}$), controlando a sensibilidade do modelo ao custo dos itens comprados de diferentes fornecedores. Valores mais altos de α priorizam a minimização do custo de aquisição dos produtos, buscando soluções que favoreçam preços mais baixos.

Por outro lado, o parâmetro β pondera o custo de realizar um pedido ao fornecedor (O_jY_{jt}). Ele ajusta a frequência de pedidos, já que um β elevado incentiva a redução do número de pedidos para minimizar os custos de transação, favorecendo compras em maior volume e menor frequência.

O terceiro componente da função objetivo é ponderado por γ, que está associado ao custo de manutenção do estoque (H_iS_{it}). Esse parâmetro regula o nível de estoque mantido ao longo do tempo, pois quanto maior o valor de γ, maior será o peso dado aos custos de armazenagem e, consequentemente, o modelo tenderá a manter estoques menores.

As restrições (4.7) atualizam os níveis de estoque nos períodos subsequentes, garantindo que os produtos entregues e a demanda atendida sejam devidamente contabilizados em cada período.

As restrições (4.8) asseguram que a quantidade de produtos entregue por um fornecedor seja condicional à ativação do respectivo pedido, ou seja, um pedido só pode ser realizado se a variável binária Y_{jt} estiver ativada. Essa restrição também garante que as entregas não excedam a demanda acumulada, impondo um controle na quantidade fornecida.

Além disso, as restrições (4.9) introduzem o conceito de estoque de segurança ponderado pelo parâmetro δ, que multiplica a soma da demanda acumulada até o período t. Isso assegura que o nível de estoque seja sempre, no mínimo, δ vezes a demanda acumulada até aquele período, garantindo um nível mínimo de segurança nos estoques para mitigar os riscos de falta de produtos.

As restrições (4.10) definem as variáveis Y_{jt} como binárias, indicando se um pedido foi feito ao fornecedor j no período t, e as restrições (4.11) impõem a não negatividade das variáveis X_{ijt}, que representam as quantidades entregues pelos fornecedores.

Dessa forma, os parâmetros α, β e γ desempenham papéis importantes no equilíbrio entre custo de compra, custo de pedido e custo de manutenção de estoque, enquanto δ ajuda a garantir níveis mínimos de estoque para segurança operacional. Juntos, eles proporcionam ao modelo flexibilidade para se ajustar a diferentes realidades operacionais e preferências de política de estoque.

4.1.4 Treinamento

O treinamento utilizando o algoritmo PPO é dividido em várias etapas cruciais, cada uma responsável por diferentes aspectos do aprendizado do agente. O objetivo principal é otimizar as decisões de reabastecimento de forma a maximizar a recompensa cumulativa, que é diretamente influenciada pela minimização dos custos operacionais de estoque e pela prevenção de rupturas ou excessos. O processo de treinamento dos aspectos técnicos incluem os seguintes itens:

1. **Coleta de Experiência:**

    - Durante a simulação do ambiente, o agente PPO interage com o sistema de gerenciamento de estoque ao longo de um horizonte temporal, tomando decisões baseadas no estado atual do sistema.
    - A cada interação, são registrados pares *estado-ação-recompensa*, onde o estado captura informações como níveis de estoque e demanda, a ação representa as decisões de reabastecimento, e a recompensa reflete o impacto dessas decisões nos custos operacionais. Esse processo é repetido até que um número suficiente de amostras seja coletado para atualizar a política do agente.

2. **Atualização da Política:**

    - Após a coleta de experiência, o PPO utiliza essas amostras para atualizar a política. A política determina a probabilidade de escolher uma ação dado um estado e é representada por uma função parametrizada por uma rede neural.
    - O algoritmo PPO se baseia em uma abordagem de gradiente, com uma técnica de otimização restrita. Especificamente, ele introduz uma restrição na magnitude das atualizações de política, limitando o quanto a política pode ser ajustada em cada passo por meio de uma técnica chamada *clipping*. Isso evita que o modelo faça grandes mudanças abruptas que possam desestabilizar o treinamento, garantindo um aprendizado mais estável.

3. **Ajuste dos Parâmetros:**

    - Durante o treinamento, os parâmetros da política (pesos da rede neural) e os parâmetros de controle do modelo de inventário são ajustados simultaneamente. O agente tenta otimizar esses parâmetros de forma a balancear os diferentes componentes de custo, maximizando a recompensa cumulativa.
    - A função de valor (*value function*) também é atualizada, fornecendo uma estimativa de quão boa é a política atual ao prever a recompensa acumulada futura a partir de um dado estado. Essa função auxilia na atualização da política ao calcular a diferença entre a recompensa esperada e a observada, ajustando a política para favorecer ações que maximizam essa diferença.

4.1.4.1 Avaliação e Iteração

Ao final de cada ciclo de atualização, a política é testada em novos episódios de simulação. Os resultados são comparados com as iterações anteriores para avaliar se houve melhorias, como uma redução nos custos totais ou maior eficiência no controle de estoque. O processo de treinamento continua até que a política atinja um desempenho estável, o que geralmente ocorre quando as recompensas começam a convergir.

4.1.4.2 Parâmetros do PPO e Configuração da Rede Neural

O sucesso do treinamento com PPO depende da escolha apropriada dos hiperparâmetros do algoritmo. Entre os mais importantes, destacam-se:

- **n_steps:** Número de passos de simulação antes de uma atualização da política. Define o tamanho da janela de experiência coletada a cada iteração.
- **epochs:** Número de vezes que a política será atualizada usando o conjunto de dados coletado.
- **n_envs:** Número de instâncias paralelas do ambiente de simulação rodando simultaneamente para coletar experiências, acelerando o treinamento.
- **clip_range:** Intervalo de clipping que limita a magnitude das atualizações de política, garantindo estabilidade.
- **learning_rate:** Taxa de aprendizado da política e da função de valor, controlando o tamanho das atualizações nos parâmetros da rede neural.
- **seed:** Semente utilizada para inicializar os processos de treinamento, assegurando reprodutibilidade dos resultados.

Neste trabalho, utilizamos a **MlpPolicy**, que é uma política baseada em redes neurais totalmente conectadas (*fully connected*), adequada para lidar com o ambiente de controle de estoque, onde os estados são representados por variáveis contínuas. A *MlpPolicy* é projetada para modelar diretamente a relação entre o estado do ambiente e as ações, utilizando camadas densas de neurônios. Essa estrutura é ajustada para extrair as características mais relevantes do ambiente, permitindo que o modelo tome decisões eficientes de reabastecimento ao gerar as ações apropriadas com base nos estados observados, otimizando assim o processo de gerenciamento de estoque.

A arquitetura da rede é formada por dois componentes principais:

- **Ator:** Responsável por selecionar as ações a serem tomadas com base nas observações do ambiente. O ator aprende a maximizar a recompensa esperada ao longo do tempo.

- **Crítico:** Avalia a qualidade das ações escolhidas pelo ator, estimando o valor esperado das observações do ambiente. O crítico serve de base para ajustar a política, orientando o ator a tomar decisões melhores.

Ambos, ator e crítico, compartilham a primeira parte da rede, chamada extrator de características, que processa as observações do ambiente e extrai informações relevantes para as decisões. Na segunda parte, redes totalmente conectadas (*fully connected*) processam as características extraídas para gerar as ações (no caso do ator) ou as estimativas de valor (no caso do crítico).

Foi utilizado a função de ativação **tangente hiperbólica**, que garante que os valores intermediários entre os nós da rede neural sejam mantidos dentro de um intervalo limitado, o que ajuda a estabilizar o treinamento, especialmente em ambientes com grande variabilidade de observações e recompensas.

A arquitetura da rede neural é controlada pelos parâmetros *net_arch* e *features_extractor*, que definem, respectivamente, a profundidade e o número de neurônios da rede, assim como o tipo de extrator de características utilizado. Essas definições permitem ajustar a rede à complexidade do ambiente, garantindo que o modelo possa aprender a lidar de forma eficiente com os desafios da gestão de inventário. A Figura 4 ilustra a separação entre as redes do ator e do crítico no PPO.

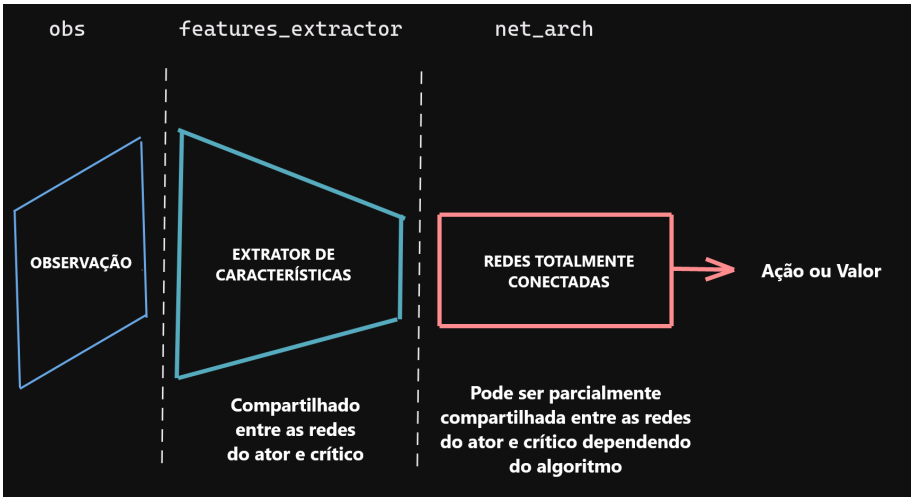

Figura 4 – Estrutura do Ator e Crítico no PPO. Adaptada (RAFFIN et al., 2021)

Dessa forma, a configuração da rede neural e os hiperparâmetros do PPO são fundamentais para garantir a eficiência do aprendizado e a obtenção de políticas de reabastecimento otimizadas.

5 Resultados

No capítulo de resultados, é realizada uma comparação entre as políticas determinísticas e as políticas treinadas utilizando o algoritmo PPO. O objetivo é avaliar como o aprendizado por reforço impacta o desempenho das decisões de reabastecimento de estoque ao longo do tempo, especialmente em termos de minimização de custos e gestão de estoque. Ao longo dos experimentos, foi analisado a evolução das recompensas acumuladas, comparando a eficiência das políticas treinadas em relação às políticas determinísticas previamente definidas.

5.1 Instâncias

As instâncias utilizadas para o treinamento e simulações envolvem um cenário de gerenciamento de estoque com três produtos ($I = 3$) e dois fornecedores ($J = 2$). O ambiente foi simulado por um horizonte de $T = 20$ passos ou períodos, onde a demanda (D) para cada produto em cada período foi gerada aleatoriamente com valores inteiros entre 1 e 200. O preço (P) de cada produto para cada fornecedor também foi definido de forma aleatória, variando entre 5 e 10. O custo de manutenção (H) foi considerado entre 0 e 3 por período para cada produto, enquanto o custo de pedido (O) variou entre 51 e 80 para cada fornecedor. Adicionalmente, o custo de falta (F) foi gerado aleatoriamente com valores entre 15 e 20 para cada produto, e o custo fixo de falta (f_F) variou entre 60 e 100.

5.2 Políticas Determinísticas

As combinações de parâmetros α, β, γ e δ foram escolhidas para representar diferentes prioridades no processo de otimização do estoque. A combinação $(1,0,0,0)$ enfatiza o custo de manutenção, enquanto $(0,1,0,0)$ prioriza o custo de pedido, e $(0,0,1,0)$ foca no custo de falta. A política $(1,1,1,0)$ busca equilibrar todos os três componentes de custo sem penalização de estoque de segurança, e $(1,1,1,0.5)$ adiciona uma leve penalização para o estoque de segurança. Essas combinações permitem avaliar como o modelo responde a diferentes configurações de prioridades de custo e seu impacto no desempenho geral. Para cada uma dessas políticas, foi realizada uma análise estatística baseada nas recompensas acumuladas ao longo dos episódios simulados. As métricas de desempenho incluem a média das recompensas, média dos parâmetros, intervalos de confiança, desvio padrão e a decomposição dos custos operacionais. Os resultados serão apresentados posteriormente em gráficos comparando os resultados com a política treinada. Essas instâncias foram escolhidas e adaptadas durante o processo de implementação com o objetivo de testar o modelo em um problema de pequena escala, facilitando a observação do comportamento do treinamento.

5.3 Treinamento

Nesta seção, é apresentado os resultados do treinamento do ambiente de simulação utilizando o algoritmo PPO para a busca da melhor combinação de políticas que controlam as decisões de reabastecimento no modelo de inventário. A combinação dos parâmetros de decisão foi ajustada dinamicamente pelo PPO para maximizar a recompensa acumulada, representada pela minimização dos custos operacionais de estoque e pela prevenção de rupturas ou excessos. O tempo total de treinamento das políticas para 1 milhão de passos foi por volta de 1 hora e 40 minutos. Sendo que cada passo (*step*) durou menos de 1 segundo.

5.3.1 Curva de Aprendizado

O ambiente de simulação foi configurado para ajustar iterativamente os parâmetros do modelo de inventário, com o objetivo de maximizar a eficiência das decisões de reabastecimento. A Figura 5 ilustra a curva de aprendizado obtida durante o treinamento com PPO. Esta curva representa a evolução da recompensa média ao longo dos episódios de treinamento, mostrando a convergência gradual das políticas otimizadas.

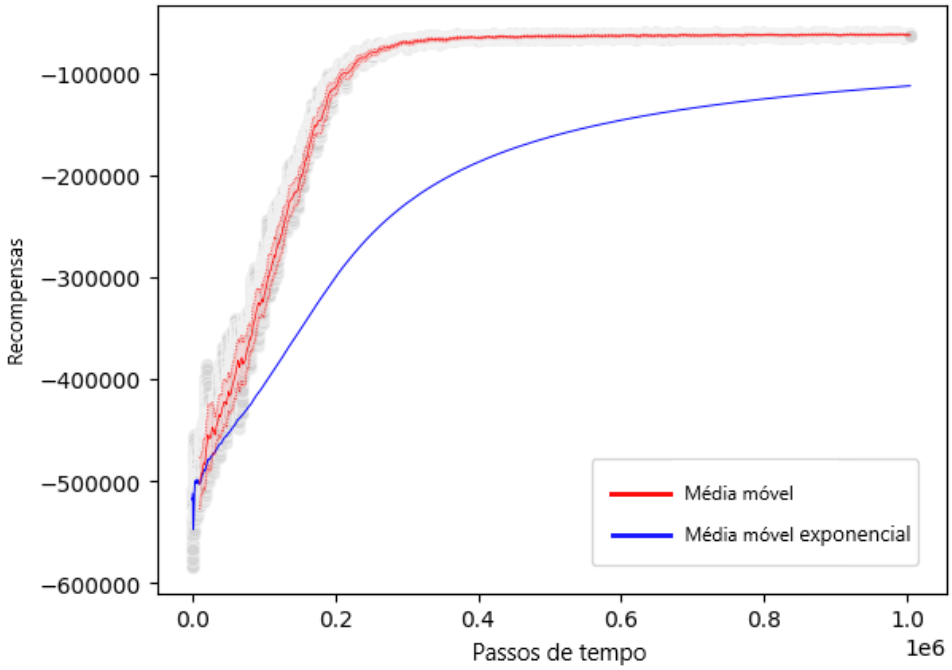

Figura 5 – Curva de aprendizado durante o treinamento com PPO.

A curva de aprendizado foi construída com base em diferentes elementos que permitem analisar o comportamento do modelo durante o treinamento:

- **Recompensas em cada interação:** Os pontos cinza claros representam as recompensas obtidas em cada interação do modelo com o ambiente ao longo dos *timesteps*, com um grau de transparência ($\alpha = 0.5$) para ilustrar a variabilidade inicial nas recompensas durante o treinamento. Esta dispersão é esperada nas primeiras fases, devido à natureza exploratória do algoritmo PPO.

- **Recompensa esperada (linha azul):** A linha azul indica a *exponential moving average* (ExRewards) das recompensas, que suaviza as flutuações imediatas das recompensas obtidas a cada *timestep*. Isso oferece uma visão mais clara da tendência geral de melhoria do desempenho do agente ao longo do tempo, à medida que ele ajusta suas políticas.

- **Média móvel (linha vermelha):** A linha vermelha representa a média móvel (*MaRewards*) das recompensas, que serve como um indicador do progresso do treinamento. Esta linha tende a suavizar as variações de curto prazo, destacando a direção geral de convergência das políticas do modelo.

- **Intervalo de confiança (linhas tracejadas vermelhas):** As linhas vermelhas tracejadas delimitam o intervalo de confiança da média móvel, representando o limite superior (*UbMaRewards*) e o limite inferior (*LbMaRewards*). Esses limites indicam a incerteza sobre a média móvel e ajudam a avaliar a estabilidade do treinamento.

- **Área sombreada:** A área sombreada em vermelho clara entre as linhas tracejadas indica visualmente o intervalo de confiança, destacando a faixa na qual as recompensas médias podem variar. À medida que o treinamento avança, essa área tende a se reduzir, sugerindo uma maior estabilidade no aprendizado.

É importante destacar que, no contexto deste modelo de inventário, a recompensa é negativa porque ela está diretamente associada aos custos operacionais totais do sistema. Ou seja, a função de recompensa foi definida como a soma de penalidades relacionadas aos custos de pedido, manutenção de estoque, e falta de produtos. Assim, à medida que o agente aprende a tomar decisões que minimizam esses custos, a recompensa negativa diminui em magnitude, indicando uma melhoria na performance do sistema. O objetivo final do treinamento é, portanto, minimizar essa recompensa negativa, o que equivale à maximização da eficiência operacional e à redução dos custos no longo prazo.

No início do treinamento, há uma grande variação nas recompensas obtidas, refletindo a fase exploratória inicial, onde o agente está experimentando diferentes ações para aprender sobre o ambiente. Com o tempo, as recompensas convergem para valores mais altos e consistentes, à medida que o agente refina suas políticas de reabastecimento. A curva de aprendizado mostra claramente que, após um número significativo de interações com o ambiente, o modelo PPO começa a estabilizar suas decisões, alcançando uma convergência gradual, o que indica que o agente aprendeu a tomar decisões mais eficazes com base nas observações do ambiente.

5.4 Comparação com Políticas Determinísticas

Para avaliar o impacto do aprendizado por reforço sobre as decisões de reabastecimento, compara-se o desempenho das políticas treinadas pelo PPO com políticas determinísticas predefinidas. Essas políticas determinísticas utilizam valores fixos para as decisões de reabastecimento, sem a adaptação dinâmica oferecida pelo aprendizado.

A Tabela 3 resume os resultados obtidos nas simulações, apresentando as recompensas médias, intervalo de confiança, variância e desvio padrão das recompensas para cada configuração de política com α, β, γ e δ determinística, além da política treinada com PPO. As simulações mostram que o PPO foi capaz de identificar combinações de decisões que reduziram significativamente os custos operacionais em comparação com as políticas determinísticas.

Tabela 3 – Resultados das Simulações para as Políticas Determinísticas e a Política Treinada com PPO

Configuração	**Média da Recompensa e Intervalo de Confiança**	**Desvio Padrão**
$[1,0,0,0]$	-73783.96 ± 974.13	4995.11
$[0,1,0,0]$	-75053.04 ± 1070.17	5487.57
$[0,0,1,0]$	-75909.31 ± 1352.86	6937.11
$[1,1,1,0]$	-69479.76 ± 1365.64	7002.66
$[1,1,1,0.5]$	-66719.50 ± 841.97	4317.39
Política Treinada	-56807.44 ± 993.62	5095.01

Tabela 4 – Valores Médios e desvio padrão para α, β, γ e δ da política treinada com PPO

Política Treinada	**Alpha**	**Beta**	**Gamma**	**Delta**
Valores médios	1.00	0.50	1.00	0.61
Desvio padrão	0.00	0.03	0.00	0.05

O desvio padrão reportado na Tabela 3 reflete a variabilidade das recompensas observadas durante as simulações. Um desvio padrão mais elevado, como observado nas políticas determinísticas $[0,0,1,0]$ e $[1,1,1,0]$, indica uma maior dispersão nos custos operacionais ao longo dos episódios simulados, o que pode ser associado a uma maior incerteza no desempenho dessas configurações.

A Tabela 4 apresenta os valores médios e o desvio padrão para os parâmetros α, β, γ e δ da política treinada com o algoritmo PPO. Os resultados mostram que α e γ mantiveram o valor médio de 1,0, com desvio padrão de 0,0, indicando que esses parâmetros não foram alterados ao longo do treinamento, devido ao impacto positivo na otimização dos custos associados. Já o parâmetro β apresentou um valor médio de 0,5, com um desvio padrão de 0,03, o que demonstra um impacto relacionado ao custo de pedido com uma leve variação durante o processo de treinamento. O parâmetro δ, por sua vez, teve um valor médio de 0,61, com desvio padrão de 0,05, sugerindo uma variação moderada, refletindo o impacto da penalização relacionada ao estoque de segurança. Esses resultados indicam que a política treinada foi capaz de ajustar α, β, γ e δ de maneira eficaz, com pouca variabilidade nas decisões associadas a α e γ, enquanto β e δ apresentaram uma maior flexibilidade nas ações ao longo do tempo.

É importante destacar que, ao contrário das políticas determinísticas, os valores dos parâmetros da política treinada pelo PPO são ajustados dinamicamente a cada estado observado no ambiente. Isso significa que esses parâmetros não são fixos durante todo o processo de simulação, mas variam de acordo com a condição do sistema de inventário, buscando otimizar a decisão de reabastecimento em função das características atuais do ambiente, como o nível de estoque, a demanda esperada e os custos operacionais. Essa capacidade adaptativa contribui para a superioridade da política treinada em termos de minimização dos custos e eficiência na gestão de estoque.

A Figura 6 também apresenta um gráfico *boxplot* que compara a distribuição das recompensas obtidas por ambas as abordagens. Pode-se ver que a política treinada com PPO obteve menores custos médios e maior consistência, evidenciada pela menor variabilidade das recompensas.

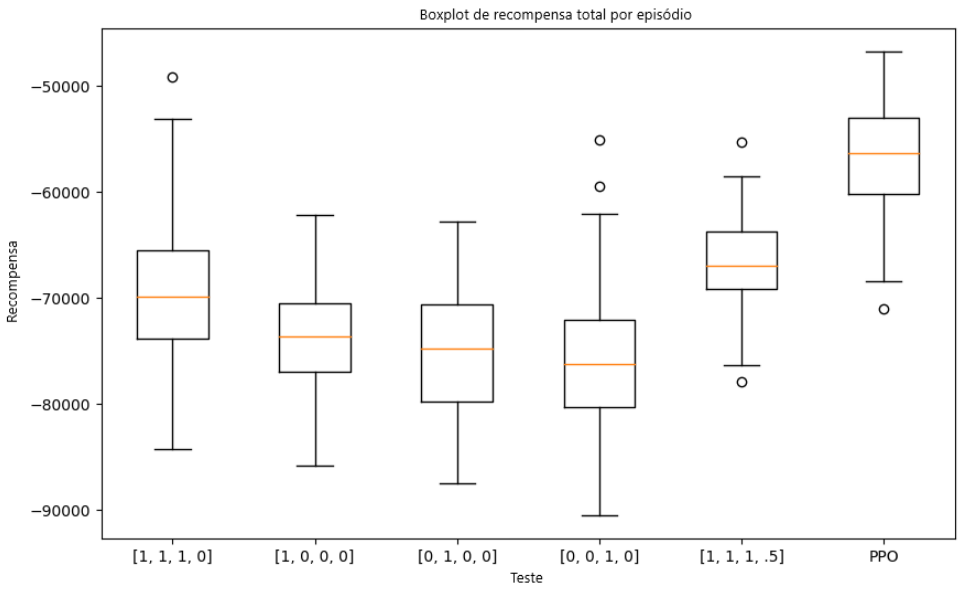

Figura 6 – Comparação das distribuições de recompensa entre as políticas treinadas e determinísticas.

Além do *boxplot*, o gráfico apresentado na Figura 7 apresenta a distribuição dos custos operacionais médios para cada uma das políticas. Esse gráfico de barras ilustra de forma clara como a política treinada com PPO resultou em menores custos totais em comparação com as políticas determinísticas, evidenciando a eficácia do aprendizado por reforço.

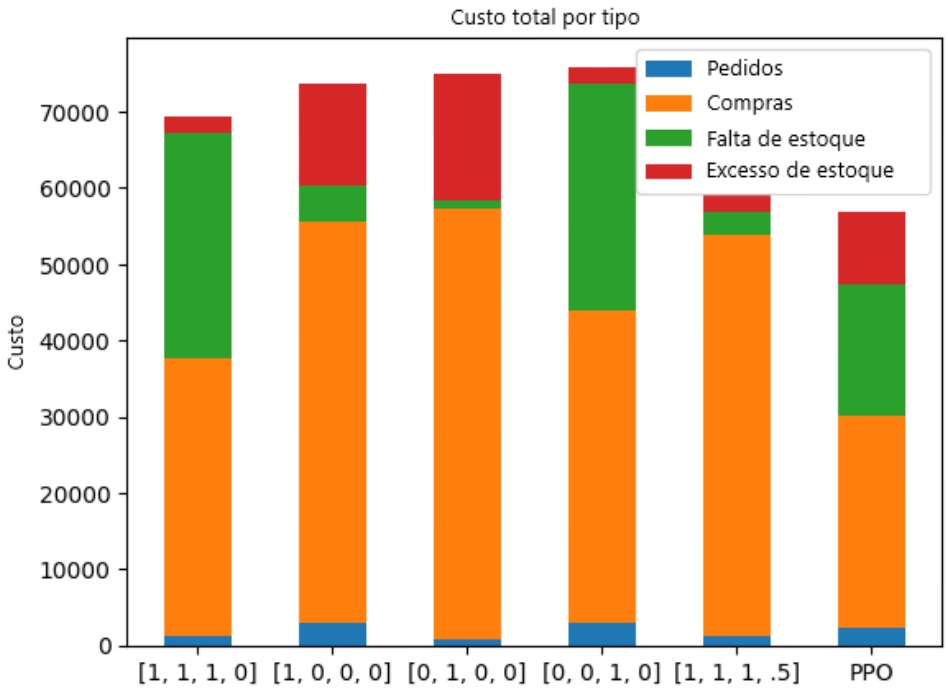

Figura 7 – Distribuição dos custos operacionais médios entre políticas treinadas e determinísticas.

Como pode-se observar, a política PPO obteve menores custos operacionais médios em relação a todas as outras políticas determinísticas analisadas. A diferença mais significativa ocorre quando se compara a política treinada com a configuração $[0,0,1,0]$, que apresentou o maior custo entre as políticas determinísticas. Esse resultado reflete a capacidade do modelo PPO de ajustar os parâmetros de maneira eficiente em um ambiente de demanda estocástica, otimizando o trade-off entre custos de manutenção, pedidos e falta de estoque.

O gráfico da Figura 8 mostra a distribuição de densidade das recompensas obtidas em cada política. Essa visualização permite uma análise mais detalhada da dispersão das recompensas em torno da média, destacando as políticas que tiveram maior consistência no desempenho.

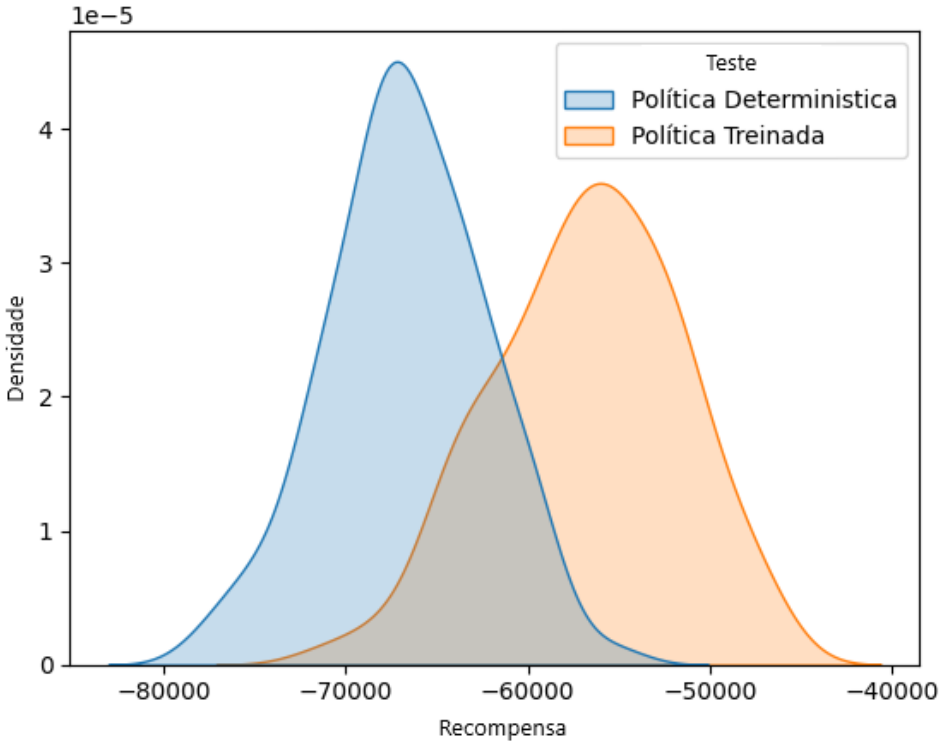

Figura 8 – Distribuição de densidade das recompensas entre políticas treinadas e determinísticas.

A curva de densidade da política PPO apresenta um pico mais concentrado em torno de recompensas mais elevadas, o que reflete sua maior eficiência. Em contraste, as políticas determinísticas apresentam maior dispersão, com algumas recompensas significativamente abaixo da média.

O gráfico da Figura 9 mostra as recompensas médias para cada política, com a indicação dos respectivos intervalos de confiança (erro padrão). Esse gráfico permite visualizar de forma clara as diferenças entre as recompensas obtidas pelas políticas treinadas e determinísticas.

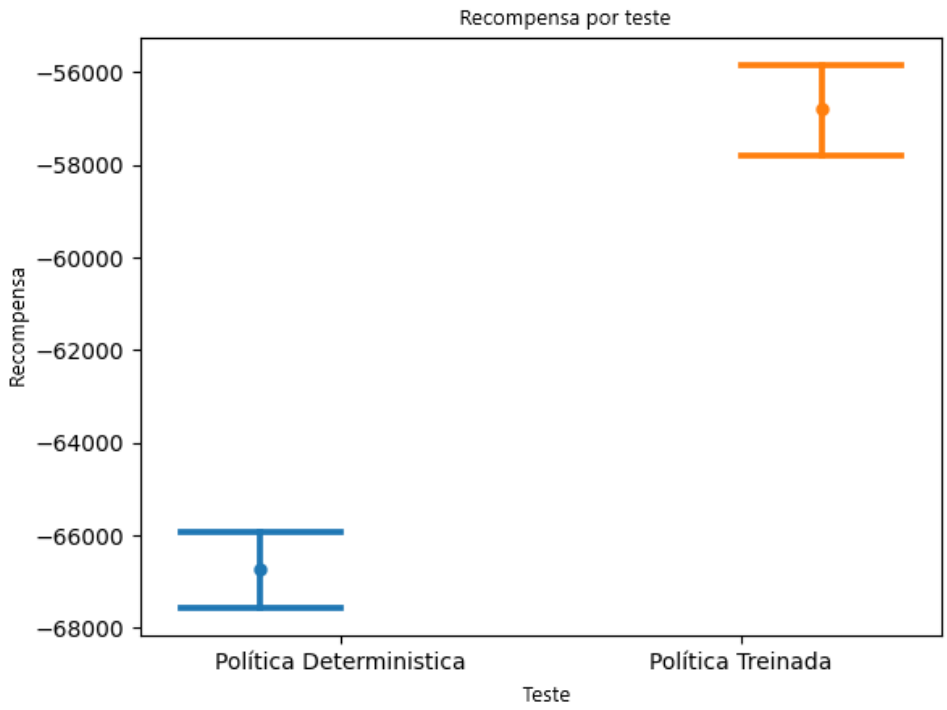

Figura 9 – Relação entre ações tomadas e recompensas obtidas.

No gráfico de pontos, as recompensas para cada política estão dispostas ao longo do eixo x, enquanto o eixo y representa o valor médio da recompensa. As barras de erro correspondem ao intervalo de confiança, oferecendo uma visão clara da variabilidade das recompensas. A política PPO apresenta recompensas médias mais altas, com menores margens de erro, o que indica um desempenho mais consistente em comparação com as políticas determinísticas. As políticas determinísticas, por outro lado, mostram recompensas mais baixas e com maior dispersão, refletindo um desempenho menos confiável.

Para avaliar a eficácia da política treinada utilizando o PPO, foi realizada uma comparação estatística entre os resultados obtidos pela política treinada e a melhor política determinística encontrada. A hipótese nula (H_0) testada nesse contexto é que não há diferença significativa entre as recompensas médias obtidas pela política treinada com PPO e pela política determinística $[1,1,1,0.5]$, enquanto a hipótese alternativa (H_1) propõe que há uma diferença significativa entre os dois conjuntos de recompensas.

O teste estatístico utilizado foi o **teste t pareado**, implementado através da função *ttest_rel* da biblioteca *SciPy*, que compara diretamente as recompensas de ambas as políticas, assumindo que os dados seguem uma distribuição aproximadamente normal. A estatística t resultante foi de -23.54, indicando uma diferença considerável entre as duas políticas. O valor p associado ao teste foi de 2.42×10^{-42}, que é extremamente pequeno, rejeitando com confiança a hipótese nula e indicando que a política treinada com PPO supera a política determinística $[1,1,1,0.5]$.

Esse resultado confirma que a política aprendida através do PPO apresenta desempenho superior em termos de maximização da recompensa acumulada, validando o uso do aprendizado por reforço para o problema de gerenciamento de estoque de medicamentos.

5.4.1 Análise dos Resultados

Os resultados demonstram que o uso do algoritmo PPO levou a uma política mais eficaz, que conseguiu otimizar as decisões de reabastecimento em um cenário de gestão de estoque hospitalar. Enquanto as políticas determinísticas apresentaram um desempenho fixo e limitado, a política treinada com PPO adaptou-se às condições variáveis do ambiente, resultando em menores custos operacionais e maior eficiência no controle de estoque.

Portanto, o aprendizado por reforço mostrou-se uma abordagem promissora para problemas de gestão de inventário, especialmente em cenários onde as demandas são voláteis e o controle sobre os custos operacionais é crucial para a operação. Esses resultados reforçam a importância de métodos adaptativos para a tomada de decisão em ambientes complexos.

6 Considerações Finais

Neste trabalho, foi desenvolvido um ambiente de simulação baseado em aprendizado por reforço, utilizando o algoritmo PPO, para otimizar as decisões de reabastecimento em um modelo de inventário. O sistema foi configurado para ajustar dinamicamente os parâmetros α, β, γ e δ, que afetam diretamente as políticas de controle de estoque, levando em consideração aspectos operacionais, como custo de manutenção, rupturas e excessos de estoque. A integração entre o ambiente de simulação e o *solver* Gurobi permitiu a aplicação da otimização via Programação Linear Inteira Mista (MIP) em cada etapa de decisão, possibilitando uma modelagem adaptativa do problema de inventário.

Os resultados indicam que o uso do PPO resultou em uma melhoria no desempenho das políticas de reabastecimento, quando comparadas às políticas determinísticas predefinidas. A política treinada conseguiu identificar combinações de parâmetros que minimizam os custos operacionais e contribuem para uma maior eficiência no gerenciamento de estoque.

Apesar dos progressos, o trabalho revelou a necessidade de mais testes e ajustes. O impacto de diferentes configurações de hiperparâmetros no desempenho final do modelo ainda precisa ser explorado, especialmente em cenários de maior complexidade e variabilidade de demanda. Além disso, uma análise mais detalhada da robustez do modelo diante de mudanças nos parâmetros de custo e na estrutura do ambiente pode apontar novas oportunidades de aprimoramento.

6.1 Trabalhos Futuros

Como direções para trabalhos futuros, propõe-se o desenvolvimento de novas formas de parametrização dos fatores que influenciam as decisões de reabastecimento. Uma abordagem seria explorar a introdução de novos parâmetros ou a combinação de diferentes métodos de otimização, como a inclusão de técnicas de ajuste fino nos custos de transporte e armazenamento.

Outra linha de investigação seria testar diferentes arquiteturas para a rede neural do modelo PPO, especialmente na estrutura *ator-crítico*. Alterar o número de camadas, a função de ativação ou mesmo o uso de técnicas de regularização pode aumentar a capacidade de generalização do modelo e melhorar os resultados. Também seria relevante explorar o uso de outros algoritmos de aprendizado por reforço, como o *Deep Deterministic Policy Gradient* (DDPG) ou o *Twin Delayed Deep Deterministic Policy Gradient* (TD3), para comparar o desempenho com o PPO no contexto de inventário.

Por fim, futuras pesquisas podem focar na adaptação do modelo para diferentes ambientes e cenários de inventário, além de considerar a aplicação de outras restrições e parâmetros que simulem o gerenciamento de estoques em cadeias de suprimentos mais complexas e com múltiplos produtos.

Referências

ANUPINDI, R.; TAYUR, S. Managing stochastic multiproduct systems: model, measures, and analysis. **Operations Research**, INFORMS, v. 46, n. 3-supplement-3, p. S98–S111, 1998.

BASNET, C.; LEUNG, J. M. Inventory lot-sizing with supplier selection. **Computers & Operations Research**, Elsevier, v. 32, n. 1, p. 1–14, 2005.

BERTANI, T. M. **Lean Healthcare: Recomendações para implantações dos conceitos de produção enxuta em ambientes hospitalares**. Tese (Doutorado) — Universidade de São Paulo, 2012.

CAI, Q.; HANG, W.; MIRHOSEINI, A.; TUCKER, G.; WANG, J.; WEI, W. Reinforcement learning driven heuristic optimization. **arXiv preprint arXiv:1906.06639**, 2019.

CARDOEN, B.; DEMEULEMEESTER, E.; BELIËN, J. Operating room planning and scheduling: A literature review. **European journal of operational research**, Elsevier, v. 201, n. 3, p. 921–932, 2010.

Farama Foundation. **Gymnasium Documentation**. 2023. Acesso em 04/10/2024. Disponível em: <https://gymnasium.farama.org/>.

FEIRING, B. R.; SASTRI, T. Improving production planning by utilizing stochastic programming. **Computers & Industrial Engineering**, Elsevier, v. 19, n. 1-4, p. 53–56, 1990.

FRANSOO, J. C.; SRIDHARAN, V.; BERTRAND, J. W. M. A hierarchical approach for capacity coordination in multiple products single-machine production systems with stationary stochastic demands. **European Journal of Operational Research**, Elsevier, v. 86, n. 1, p. 57–72, 1995.

GONÇALVES, A. A.; NOVAES, M. L. d. O.; SIMONETTI, V. M. M. Otimização de farmácias hospitalares: eficácia da utilização de indicadores para gestão de estoques. **XXVI Encontro Nacional de Engenharia de Produção, Fortaleza–CE**, 2006.

Gurobi Optimization, LLC. **Gurobi Optimizer Reference Manual**. 2023. Acesso em 04/10/2024. Disponível em: <https://www.gurobi.com/documentation/>.

HARRIS, C. R.; MILLMAN, K. J.; WALT, S. J. van der; GOMMERS, R.; VIRTANEN, P.; COURNAPEAU, D.; WIESER, E.; TAYLOR, J.; BERG, S.; SMITH, N. J.; KERN, R.; PICUS, M.; HOYER, S.; KERKWIJK, M. H. van; BRETT, M.; HALDANE, A.; RíO, J. F. D.; WIEBE, M.; PETERSON, P.; GÉRARD-MARCHANT, P.; SHEPPARD, K.; REDDY, T.; WECKESSER, W.; ABBASI, H.; GOHLKE, C.; OLIPHANT, T. E. **Array programming with NumPy**. 2020. 357–362 p. Acesso em 04/10/2024. Disponível em: <https://numpy.org/doc/>.

HEZEWIJK, L. V.; DELLAERT, N.; WOENSEL, T. V.; GADEMANN, N. Using the proximal policy optimisation algorithm for solving the stochastic capacitated lot sizing problem. **International Journal of Production Research**, Taylor & Francis, v. 61, n. 6, p. 1955–1978, 2023.

HUNTER, J. D. **Matplotlib: A 2D Graphics Environment**. 2007. 90–95 p. Acesso em 04/10/2024. Disponível em: <https://matplotlib.org/stable/users/index.html>.

IIDA, T. The infinite horizon non-stationary stochastic inventory problem: Near myopic policies and weak ergodicity. **European Journal of Operational Research**, Elsevier, v. 116, n. 2, p. 405–422, 1999.

KARMARKAR, U. S.; YOO, J. The stochastic dynamic product cycling problem. **European Journal of Operational Research**, Elsevier, v. 73, n. 2, p. 360–373, 1994.

KRITCHANCHAI, D.; MEESAMUT, W. Developing inventory management in hospital. **International Journal of Supply Chain Management**, v. 4, n. 2, p. 11–19, 2015.

MARTEL, A.; DIABY, M.; BOCTOR, F. Multiple items procurement under stochastic nonstationary demands. **European Journal of Operational Research**, Elsevier, v. 87, n. 1, p. 74–92, 1995.

OLIVEIRA, R. V. de; SOUZA, M. C. de; SILVA, T. Dimensionamento de lotes com seleção de fornecedores e possibilidades de desconto no custo fixo. **Anais do Simpósio Brasileiro de Pesquisa Operacional**, v. 51, 2019.

POWELL, W. B. **Reinforcement learning and stochastic optimization**. [S.l.]: John Wiley & Sons Hoboken, NJ, 2021.

PyTorch Developers. **PyTorch Documentation**. 2023. Acesso em 04/10/2024. Disponível em: <https://pytorch.org/docs/>.

RAFFIN, A.; HILL, A.; GLEAVE, A.; KANERVISTO, A.; DORMANN, N. **Stable-Baselines3: Reliable Reinforcement Learning Implementations**. 2021. Acesso em 04/10/2024. Disponível em: <https://stable-baselines3.readthedocs.io/>.

RAIS, A.; VIANA, A. Operations research in healthcare: a survey. **International transactions in operational research**, Wiley Online Library, v. 18, n. 1, p. 1–31, 2011.

RODRIGUES, W. C. **Metodologia científica**. 2007. 2 p.

SCHULMAN, J.; WOLSKI, F.; DHARIWAL, P.; RADFORD, A.; KLIMOV, O. Proximal policy optimization algorithms. **arXiv preprint arXiv:1707.06347**, 2017.

SELAU, L. P. R.; PEDÓ, M. G. B.; SENFF, D. d. S.; SAURIN, T. A. Produção enxuta no setor de serviços: caso do hospital de clínicas de porto alegre-hcpa. **Revista Gestão Industrial [recurso eletrônico].**, 2009.

SIRASKAR, R. Reinforcement learning for control of valves. **Machine Learning with Applications**, Elsevier, v. 4, p. 100030, 2021.

SOBEL, M. J.; ZHANG, R. Q. Inventory policies for systems with stochastic and deterministic demand. **Operations Research**, INFORMS, v. 49, n. 1, p. 157–162, 2001.

SUTTON, R. S.; BARTO, A. G. Reinforcement learning: An introduction. **Artificial Intelligence**, 1998.

TEAM, T. P. D. **Pandas Documentation**. 2023. Acesso em 04/10/2024. Disponível em: <https://pandas.pydata.org/docs/>.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; WALT, S. J. V.; BRETT, M.; WILSON, J.; MILLMAN, K. J.; MAYOROV, N.; NELSON, A. R. J.; JONES, E.; KERN, R.; LARSON, E.; CAREY, C.; POLAT, I. P.; FENG, Y.; MOORE, E. W.; PLAS, J. V.; LAXALDE, D.; PERKTOLD, J.; CIMRMAN, R.; HENRIKSEN, I.; QUINTERO, E. A.; HARRIS, C. R.; ARCHIBALD, A. M.; RIBEIRO, A. H.; PEDREGOSA, F.; van Mulbregt, P.; SciPy 1.0 Contributors. **SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python**. 2020. 261–272 p. Acesso em 04/10/2024. Disponível em: <https://scipy.org/doc/>.

VRIES, J. D.; HUIJSMAN, R. Supply chain management in health services: an overview. **Supply Chain Management: An International Journal**, Emerald Group Publishing Limited, 2011.

WASKOM, M.; TEAM the seaborn development. **Seaborn Documentation**. 2023. Acesso em 04/10/2024. Disponível em: <https://seaborn.pydata.org/>.

ZIPKIN, P. H. Models for design and control of stochastic, multi-item batch production systems. **Operations Research**, INFORMS, v. 34, n. 1, p. 91–104, 1986.