UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Agrícola

FERNANDO BARBOSA

Aprendizado por Reforço aplicado à Prevenção contra Fraude de Seguros Agrícolas

Reinforcement Learning applied to Fraud Prevention in Agricultural Insurance

Campinas
2023

Fernando Barbosa

Aprendizado por Reforço aplicado à Prevenção contra Fraude de Seguros Agrícolas

Dissertação apresentada ao Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Agrícola na área de Agricultura Digital.

.

Supervisor/Orientador: Stanley Robson de Medeiros Oliveira

ESTE TRABALHO CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELO ALUNO FERNANDO BARBOSA E ORIENTADA PELO PROF. DR. STANLEY ROBSON DE MEDEIROS.

Campinas

2023

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

B234a

Fernando Barbosa, 1987-
Aprendizado por reforço aplicado à prevenção contra fraudes / Fernando Barbosa. – Campinas, SP : [s.n.], 2023.

Orientador: Stanley Robson de Medeiros Oliveira.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Mitigação de risco. 2. Fraude. 3. Subsídios a agricultura. 4. Aprendizado por reforço profundo. 5. Seguro agricola. I. Oliveira, Stanley Robson de Medeiros, 1967-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. III. Título.

Informações Complementares

Título em outro idioma: Reinforcement learning applied to fraud prevention in agricultural insurance
Palavras-chave em inglês:
Risk mitigation
Fraud
Agricultural subsidies
Deep reinforcement learning
Agricultural Insurance
Área de concentração: Agricultura Digital
Titulação: Mestre em Engenharia Agrícola
Banca examinadora:
Stanley Robson de Medeiros Oliveira [Orientador]
Aryeverton Fortes de Oliveira
Jurandir Zullo Junior
Data de defesa: 20-09-2023
Programa de Pós-Graduação: Engenharia Agrícola

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0009-0003-6538-1465
- Currículo Lattes do autor: http://lattes.cnpq.br/7194079001584045

Este exemplar corresponde à redação final da **Dissertação de Mestrado** defendida por **nome por extenso**, aprovada pela Comissão Julgadora em 20 de setembro de 2023, na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.

Prof. Dr. Stanley Robson de Medeiros Oliveira – Presidente e Orientador

Prof. Dr. Jurandir Zullo Junior – Membro Titular

Dr. Aryeverton Fortes de Oliveira – Membro Titular

A Ata da defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Resumo

Contra alguns riscos associados à produção agrícola, é possível obter proteção financeira pela contratação de seguros agrícolas e, no Brasil, algumas modalidades de seguros contam com subvenção governamental de seus prêmios [1]. Entre eles, está o Programa de Seguro Rural (PSR) no qual governo, seguradoras e produtores interagem em um ambiente com diversas decisões e seus processos inerentes, como aquelas relacionadas a própria contratação da apólice, avaliação da solicitação de indenização e realização de perícias. Os diferentes vieses e ruídos presentes nessa dinâmica apontam para a necessidade de mecanismos e ferramentas que ajudem a tomada de decisão, por exemplo, na avaliação de potenciais inconsistências em uma indenização a ser paga: como realizar uma sequência de decisões seleções para uma carteira de indenização de forma a minimizar o Risco de Fraude? Neste estudo, a Prevenção contra Fraudes em Seguros Agrícolas é o tema central a partir do qual é construído o ambiente de Aprendizado por Reforço denominado *InsurEnv*, abordando o problema sob assimetria de informação a respeito da ocorrência de fraude. Ou seja, a definição de fraude é desconhecida e a inferência do Risco de Fraude é realizada com base na distribuição dos dígitos do Valor da Indenização de cada proposta analisada pelo algoritmo. Com a implementação computacional de modificações na abordagem de atuais propostas de utilização de Aprendizado por Reforço, realizados no mesmo contexto de seguros e com objetivos semelhantes, foram realizados dois experimentos: um agente aleatório, com o propósito de validar o próprio ambiente e a dinâmica de episódios; e o segundo, um agente DQN treinado com um modelo de Aprendizado por Reforço. A partir das métricas avaliadas, foi possível verificar que o ambiente permite obter agentes treinados na tarefa de construção de carteiras de indenização que de fato acumulam recompensas, provando sucesso no processo de aprendizado. Espera-se que esses modelos possam resultar na prevenção e mitigação do risco de fraude na formação de carteiras de indenizações em seguros agrícolas.

Palavras-chave: Risco agrícola, contratos fraudulentos, subvenção governamental, mitigação de fraudes, função de recompensa, *Deep learning*

[1] Preço de contratação da apólice

Abstract

In agricultural production, financial protection against various risks can be obtained through agricultural insurance. In Brazil, some types of insurance, such as the Rural Insurance Program (PSR), are partially subsidized by the government. In this multifaceted environment, multiple stakeholders—government, insurers, and producers—engage in a range of decision-making processes. These include policy contracting, indemnity claim assessments, and expert evaluations. The inherent biases and noise in this dynamic point to the need for mechanisms and tools to aid decision-making, particularly in minimizing fraud risks.This study focuses on Fraud Prevention in Agricultural Insurance and introduces a Reinforcement Learning environment called 'InsurEnv.' The environment addresses the issue under the lens of information asymmetry regarding the occurrence of fraud. In other words, the definition of fraud is unknown, and the inference of Fraud Risk is based on the digit distribution of each indemnity value analyzed by the algorithm. The study also incorporates modifications in training approaches and reward functions, drawing from similar studies in the same context.Two experiments were conducted to validate this approach: the first using a random agent to test the environment and episode dynamics, and the second using a DQN agent trained through Reinforcement Learning. The evaluated metrics confirmed that the 'InsurEnv' environment allows trained agents to accumulate rewards, thereby proving successful in the learning process. It is hoped that these models will contribute to the prevention and mitigation of fraud risk in agricultural insurance indemnity portfolios.

Keywords: Agricultural risk, fraudulent contracts, government subsidy, fraud mitigation, reward function, Deep Learning

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

UNICAMP	Universidade Estadual de Campinas
FEAGRI	Faculdade de Engenharia Agrícola
MAPA	Ministério da Agricultura, Pecuária e Abastecimento
PSR	Programa de Seguro Rural
LMI	Limite Máximo de Indenização
LMG	Limite Máximo de Garantia
Pronaf	Programa Nacional de Fortalecimento da Agricultura Familiar
CPT	*Cumulative Prospect Theory*
EUT	*Expected Utility Theory*
Ocepar	Organização das Cooperativas do Estado do Paraná
Faep	Federação da Agricultura do Estado do Paraná

Sumário

1 Introdução

Há diversos riscos no processo de produção agrícola, mas é possível mitigar a perda financeira utilizando a cobertura provida por seguros que oferecem proteção contra eventuais danos relacionadas a fenômenos naturais como seca, chuva excessiva, geadas, além das adversidades nos ambientes econômico, social e sanitário. Contudo, o preço dos prêmios para adquirir esses seguros podem dificultar a formação de mercado, por incorrer de indenizações acima do limite viável para um sistema mutualista. Nesse contexto, fraudes são de particular interesse, pois afetam a taxa de sinistros e valores de indenização no processo de renovação e, portanto, a precificação dos prêmios do ciclo subsequente.

No Brasil, há algumas modalidades de seguros rurais apoiadas pelo Governo Federal, incluindo aquelas pertencentes ao PSR (Programa de Seguro Rural), um modelo de seguro mutualista implementado por meio de lei (BRASIL, 2003). Com ele, uma subvenção governamental para contratação do seguro agrícola é disponibilizada para o custeio financeiro do prêmio para aquisição de apólice, buscando a ampliação do acesso a esse tipo de seguro entre outros benefícios sociais.

Assim, os agricultores podem firmar o contrato de seguro com uma das 14 seguradoras privadas ou autarquias participantes do PSR, passando a contar com a proteção oferecida de acordo com as respectivas coberturas. Até 2021, a importância de R$10 bilhões em indenizações já havia sido paga em contratos subvencionados pelo PSR, sendo 42% a partir de 2018 (MAPA, 2022a).

O seguro agrícola desempenha um papel fundamental na proteção de agricultores e demais envolvidos contra riscos imprevisíveis, permitindo-lhes enfrentar eventos adversos, como falhas de colheita e desastres naturais. No entanto, a ocorrência de atividades fraudulentas no setor pode comprometer sua eficácia, resultando em perdas financeiras e abalando a confiança entre os participantes. Portanto, a necessidade de mecanismos robustos e eficientes de detecção de fraudes tornou-se cada vez mais evidente.

O impacto de fraudes no sistema de seguridade, portanto, pode levar a um prejuízo para o agente responsável pela subvenção como fundos gestores. Sendo o valor do prêmio sub ou superestimado em relação a um sistema neutro de ocorrências de fraude, as subvenções são de interesse público e a governança está amplamente estabelecida em lei (BRASIL, 2003).

Mitigar os riscos de fraude das carteiras de indenização é fundamental para o funcionamento adequado do sistema de seguridade como um todo, sobretudo tendo a subvenção governamental como principal fomentador. Para isso, existem mecanismos que podem ser implementados no processo de decisão para aceitação ou rejeição de uma

proposta, por exemplo, com a utilização de modelos estatísticos ou de Aprendizado de Máquina.

Apesar de haver sucesso nesse tipo de abordagem, o fenômeno de *drift*, característico das frequentes mudanças nos métodos de fraude e condições do próprio sistema, ainda é um grande desafio na sua aplicação. Novas abordagens com base em Aprendizado por Reforço (AR) podem ser úteis, uma vez que permitem capturar a evolução das dimensões de um sistema de seguridade, como o percentual de apólices aprovadas, ao mesmo tempo em que produz a estimação de modelos probabilísticos, incluindo modelos multi-variados supervisionados, como regressões.

Para enfrentar esse desafio, o trabalho se baseia em uma revisão da literatura que abrange várias áreas interdisciplinares do conhecimento, incluindo Economia, Ciências Atuariais, Psicologia, Matemática, Estatística e Ciência da Computação. Essa revisão busca esclarecer as bases teóricas e os avanços mais recentes relacionados ao AR, detecção de fraudes e seguros agrícolas, proporcionando *insights* sobre aplicações potenciais do AR além de áreas convencionais, como a robótica.

A pesquisa não apenas contribui para avanços nas aplicações do AR, mas também possui potencial para aprimorar a credibilidade e eficácia geral das práticas de seguro, beneficiando tanto o setor agrícola quanto os demais envolvidos nessa temática.

1.1 Estrutura do Trabalho

No Capítulo 2, são apresentados o contexto de seguros agrícolas no Brasil e também os conceitos fundamentais do Aprendizado por Reforço com suas aplicações em áreas de seguros e foco no tema de fraudes. Os elementos e dinâmica que são usados na implementação do ambiente computacional são apresentados no Capítulo 3, o qual contém a proposta central deste trabalho e também são descritos os experimentos que foram utilizados para avaliar a solução sob as métricas previamente estabelecidas. Os resultados desses experimentos são discutidos no Capítulo 4 juntamente com comparativos analíticos. Finalmente, a argumentação conclusiva é apresentada no Capítulo 5.

1.2 Hipótese

A abordagem Aprendizado por Reforço possibilita a formação de carteiras de indenização que reduzem o risco de fraudes em seguros agrícolas apesar da assimetria de informação sobre a definição de fraude.

1.3 Objetivo

1.3.1 Objetivo Geral

Elaborar e implementar um ambiente de aprendizado computacional baseado em Aprendizado por Reforço, visando mitigar o risco de fraude em carteiras de indenização, e utilizando tanto dados históricos como dados gerados pela interação contínua com o ambiente.

1.3.2 Objetivos Específicos

- Modelar o problema do risco de fraude na composição de carteiras de indenização.
- Criar o ambiente de Aprendizado por Reforço e conduzir testes abrangentes para assegurar sua plena funcionalidade.

2 Revisão de Literatura

Neste capítulo, realiza-se uma revisão abrangente da literatura sobre Seguros, Aprendizado de Máquina, com ênfase em Aprendizado por Reforço, e aspectos econômicos e psico-sociais relevantes. Em seguida, serão apresentados conceitos sobre modelagem para detecção de fraudes em seguros agrícolas, com ênfase em Aprendizado por Reforço. Além disso, serão discutidas tendências, melhores práticas e abordagens para detecção e prevenção de fraudes em seleção de carteiras de indenização. Por fim, por meio de uma análise interdisciplinar, será apresentada uma base sólida para desenvolver um ambiente de treinamento robusto por Aprendizado por Reforço, capaz de lidar com complexidade e incerteza no contexto de Seguros.

2.1 O Programa de Seguro Rural (PSR)

Desde sua criação em 2003 (BRASIL, 2003), a participação do Seguro Rural no orçamento da União cresceu significativamente, atingindo 11.9% no ano de 2019. Entretanto, conforme apontado por Amaral e Bacha (2022), sua participação já tinha alcançado o patamar de 18,2% no ano de 2014, ou seja, o valor dedicado a esses programas pode variar significativamente e não é garantido que seja o suficiente para o bom funcionamento do sistema. A participação do PSR nos gastos governamentais com apoio ao setor rural foi de apenas 1,9%, no período de 2000 a 2015, quase 12 vezes menor do que o valor dedicado ao Pronaf, por exemplo (SANTOS; FREITAS, 2017).

Em 2021, aproximadamente 72% dos produtores já haviam acessado os benefícios do PSR, o que demonstra a disponibilidade de dados que permitirão uma avaliação dos contratos passados e a evolução dos produtores e das culturas do programa. No mesmo ano, o programa atendeu a mais de 100 mil agricultores com uma área total de aproximadamente 14 milhões de hectares, e ultrapassando R$ 68 bilhões segurados (MAPA, 2021).

O Comitê Gestor Interministerial do Seguro Rural é o agente responsável pelo gerenciamento das subvenções. No ano de 2023, o programa disponibiliza um orçamento de mais de R$1 bi, sendo a subvenção de até 40% do prêmio para todas as culturas participantes do programa, limitada a R$60 mil reais por grupo de atividade e segurado. Além dos requisitos básicos do programa, no âmbito do seguro agrícola, é necessário que a gestão da cultura esteja de acordo com as regras contratualmente estabelecidas, sob pena de perder o direito à indenização da importância contratada.

Os produtos de seguros podem ser formatados e operacionalizados de diversas

maneiras pelas próprias seguradoras, desde que estejam de acordo com a legislação e regulamentação vigentes. Com a aprovação de uma proposta de seguro, o produtor passa a contar com a proteção financeira de acordo com as características do produto contratado bem como sua forma: multirriscos ou riscos nomeados. Cabe ressaltar alguns aspectos importantes desse acordo: os riscos não estão cobertos pelo seguro [1] como também a aplicação das orientações do ZARC [2]. Além disso, o descumprimento de tais requisitos pode levar a anulação do compromisso da seguradora, impactando terminantemente o processo indenizatório.

Após aprovação do orçamento que será dedicado ao Programa de Seguro Rural pelo Congresso Nacional, cada seguradora pode formular e ofertar produtos com diferentes configurações, por exemplo, de cobertura. Contudo, seus produtos devem cumprir as regras estabelecidas por regulamentação e legislação vigentes, por exemplo quanto ao limite anual [3], conforme demonstrado na Tabela 1.

Tabela 1 – Limites do PSR - 2023

Grupo de Atividades	Percentual de Subvenção	Limite anual
Grãos (Soja)	20%	R$ 60.000
Grãos (Outros) Frutas, Olerídolas, Café e Cana-de-açucar Florestas Pecuária Aquicultura	40%	
Limite Anual R$120.000		

Fonte: Adaptado de (MAPA, 2022a)

Por exemplo, há seguros que oferecem coberturas para eventos específicos como a ocorrência de granizo na cultura segurada (MAPA, 2021). Outras seguradores ofertam a cobertura do custeio da safra (SEGURADORA, 2019), enquanto outros produtos são dedicados à cobertura do faturamento (BRASIL, 2022).

Ao decidir obter um seguro, o produtor submete uma proposta de contratação para a seguradora. A seguradora, por sua vez, tem o papel de avaliar a proposta, podendo ou não aceitá-la, e consequentemente transmiti-la ao Ministério da Agricultura, Pecuária e Abastecimento (MAPA). É importante notar que as seguradoras podem avaliar a proposta de seguro sem necessariamente ter que aceitar a proposta realizada pelo cliente, isto é, existem casos em as propostas podem ser diretamente rejeitadas pelas seguradoras mesmo que estejam em cumprimento com a regulamentação estabelecida pela SUSEP

[1] Riscos Excluídos na terminologia atuarial

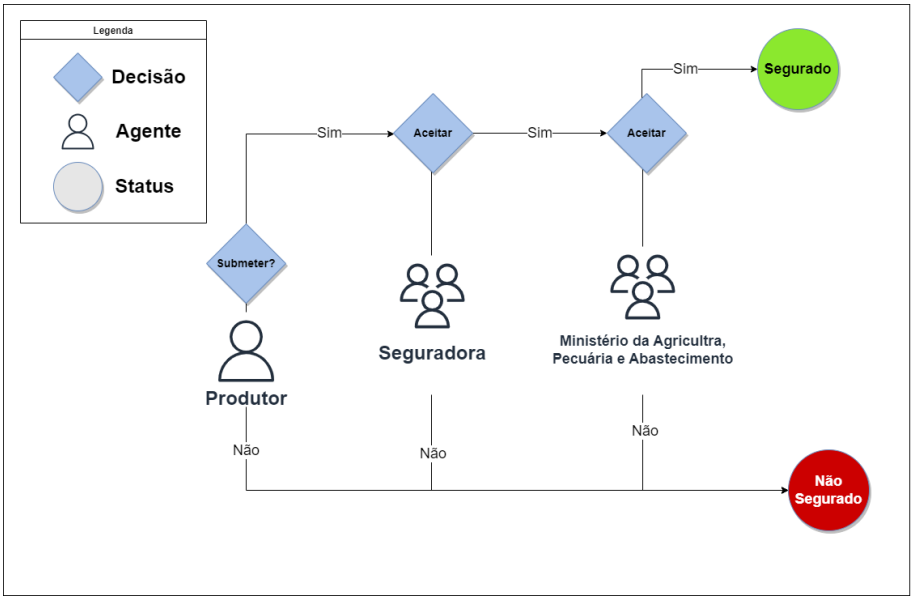
[2] Para mais informações sobre o ZARC acessar o link: <https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/programa-nacional-de-zoneamento-agricola-de-risco-climatico>

[3] Por CPF CNPJ

e de acordo com a legislação corrente. Uma seguradora pode rejeitar uma proposta de seguro (SUPERINTENDENTE DA SUPERINTENDÊNCIA DE SEGUROS PRIVADOS - SUSEP, 2004) em situações de risco elevado, informações incorretas ou omitidas, cobertura não disponível, histórico de sinistros anteriores e falta de idoneidade do proponente. É essencial que as seguradoras atuem em conformidade com as leis e regulamentações locais, garantindo uma recusa justa e transparente, baseada em critérios objetivos e razoáveis.

Na Figura 1, um exemplo do fluxo de contratação de um seguro é apresentado, abstraindo-se as especifidades de cada produto e seguradora. Primeiramente, um produtor interessado em obter um seguro rural pode considerar contratar o seguro e contar com a proteção das coberturas ou deixar de contratar o seguro e assumir os respectivos riscos.

Figura 1 – Exemplo de Processo de Contratação do Seguro.



Fonte: Elaborado pelo autor

O produtor pode optar por solicitar a indenização por eventuais perdas incorridas relativas a uma área protegida contratualmente, porém, cabe à Seguradora a decisão sobre o processo indenizatório, podendo executar auditorias e perícias mesmo antes da decisão sobre o pagamento da indenização. Contudo, determinar se uma determinada indenização é potencialmente fraudulenta é uma tarefa que pode acontecer em diversos momentos após a solicitação, inclusive nunca chegar a ser conhecida. Dessa forma, ressalta-se dois pontos importantes de decisão tomadas por parte das seguradoras: aprovar ou rejeitar a indenização. Para tanto é preciso também determinar se essa solicitação pode ser paga sem representar risco de fraude.

Caso a proposta seja aprovada, o valor da subvenção é comprometido do orçamento dedicado ao programa. No atual formato, o Departamento de Gestão de Riscos do MAPA é responsável por avaliar a solicitação de subvenção e, havendo aprovação para

utilização dos recursos, o beneficio é concedido por meio da liquidação de parte do prêmio estabelecido para o contrato.

A seguir, os valores de subvenção são repassados às seguradoras: esse momento configura de forma prática o interesse da sociedade civil na aplicação de recursos públicos. Além disso, o PSR possui condições básicas para a participação do programa, caso não sejam atendidas, o produtor não terá acesso ao programa e terá sua proposta indeferida, ou seja, não poderá contar com a subvenção do Governo para aliviar os custos do prêmio (MAPA, 2022a).

De acordo com o Art. 2o da Lei Brasil (2003), o seguro rural abrange as seguintes modalidades:

- I - seguro agrícola;
- II - seguro pecuário;
- III - seguro aquícola;
- IV - seguro de florestas;
- V - seguro de penhor rural;
- VI - seguro de benfeitorias e produtos agropecuários; e
- VII - seguro de vida.

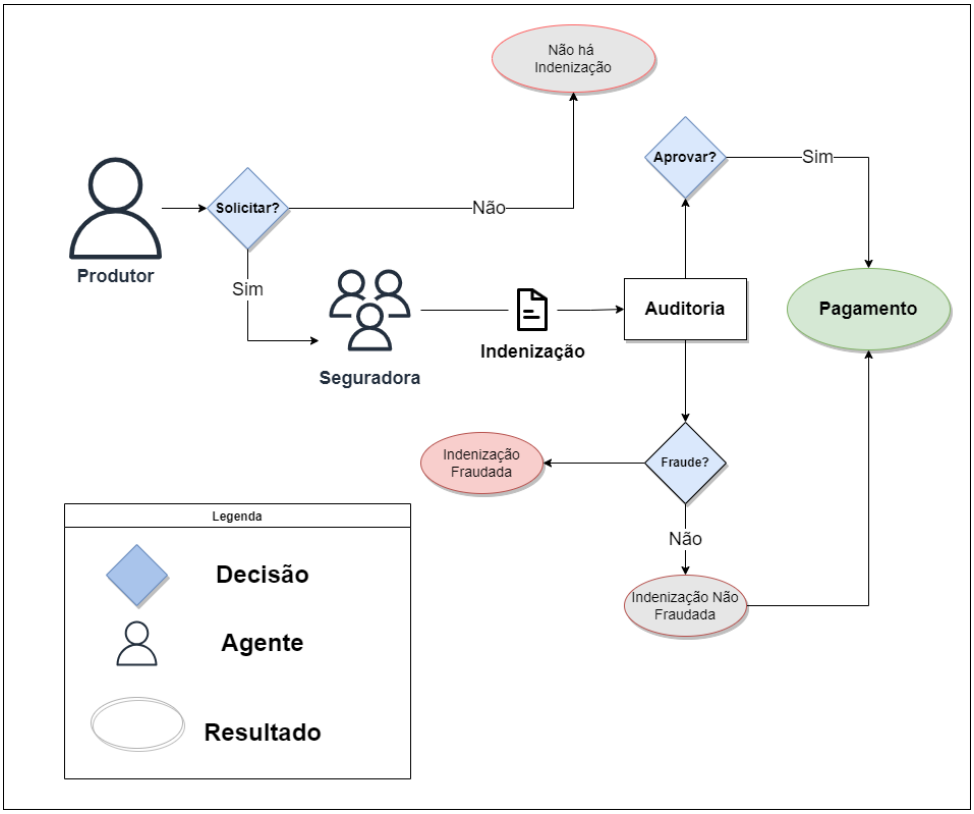
Em caso de sinistro, o produtor deve informar a seguradora com tempestividade, providenciando as informações e também permitindo acesso dos peritos ao local, de acordo com as orientações e regras estabelecidas na sua apólice de seguro (MAPA, 2022a). A avaliação das solicitações de indenização é feita pelas seguradoras e, por vezes, guiada por recomendações construídas a partir de Modelos Estatísticos ou de Inteligência Artificial (SINAYOBYE; KIWANUKA; KYANDA, 2018).

O treinamento dos peritos é regulamentado Brasil (2020) e realizado pelas próprias seguradoras, o qual possui regras e procedimentos a serem seguidos para a verificação das condições e do estado geral da lavoura nos diferentes pontos de contato durante a vigência da apólice.

Contudo, não há garantia de que o sistema do PSR como um todo implemente estratégias de redução de ruído sistêmico. A variabilidade indesejada causada por julgamentos humanos de forma sistêmica (KAHNEMAN et al., 2021) pode levar a uma distribuição inadequada do orçamento dedicado aos sinistros, com diversas consequências para o mercado.

Na Figura 2, é apresentado um exemplo genérico do processo de solicitação de indenização do seguro rural, abstraindo particularidades do processo que dependem da política e legislação vigentes.

Figura 2 – Exemplo de Processo de Auditoria.



Fonte: Elaborado pelo autor

O Guia do Seguro Rural Produzido pelo MAPA (MAPA, 2022a) destaca as práticas e requisitos para garantir o recebimento da indenização em caso de sinistro em apólices de seguro rural. É enfatizada a importância de entender e cumprir os termos acordados na contratação do seguro, como informar corretamente os detalhes do projeto de plantio, seguir o Zoneamento Agrícola de Risco Climático (ZARC) e observar os riscos excluídos. Além disso, é recomendado registrar todas as comunicações com a seguradora, cumprir os prazos estabelecidos para as vistorias e, em caso de discordância, manifestar por escrito. Além disso, as recomendações mostram que, com reclamações bem fundamentadas, é possível solicitar uma vistoria para reanálise da lavoura. Em caso de dúvidas ou reclamações, o segurado pode entrar em contato com a corretora, o SAC da seguradora ou a plataforma ***consumidor.gov.br***, bem como a Superintendência de Seguros Privados (SUSEP)

2.2 Efeitos do PSR

Para verificar se o PSR teve o efeito esperado, é importante estabelecer primeiramente quais são seus objetivos. De acordo com a SUSEP,

> O objetivo maior do Seguro Rural é oferecer coberturas que, ao mesmo tempo, atendam ao produtor e à sua produção, à sua família, à geração de garantias a seus financiadores, investidores, parceiros de negócios, todos interessados na maior diluição possível dos riscos, pela combinação dos diversos ramos de seguro.

Conforme definido pelo Tribunal de Contas da União:

> O objetivo maior do Seguro Rural é o de transferir o risco climático, o risco decorrente de problemas sanitários e os riscos de volatilidade de preços da produção agropecuária, aos quais os produtores rurais estão expostos, para outros agentes econômicos, como seguradoras e resseguradoras.

Uma terceira definição do objetivo (MAPA, 2022b):

> O objetivo do programa é garantir que o produtor sinta segurança em investir na atividade agrícola sem temer aos possíveis prejuízos e adversidades causados pelos eventos climáticos e demais perdas não gerenciáveis.

Dessa forma, cabe estabelecer se o Programa tem sido bem-sucedido tanto do ponto de vista do segurado, no aspecto financeiro e microeconômico, bem como no âmbito da oferta, por meio indicadores macroeconômicos da oferta.

A partir de variáveis relacionadas à produção e contratos de seguro do PSR, no estudo realizado por Tabosa e Filho (2021), foi proposto um modelo de vetor autorregressivo (VAR) espacial que objetiva verificar se o seguro rural tem o efeito desejado de manutenção do investimento e condições concorrenciais. Assim, foram obtidas variáveis de área cultivada, produtividade, número de apólices e agricultores bem como a *importância segurada* disponibilizadas pelo MAPA. É importante ressaltar que a relação causal entre as variáveis é presumida no estudo e os autores buscaram traduzi-la na especificação do modelo, sendo o período analisado de 2016 a 2017, para todas as culturas participantes do programa no período.

Tabosa e Filho (2021) propuseram dois modelos: o primeiro considera todas regiões brasileiras, exceto Roraima, Amapá e o Distrito Federal; o segundo, considera as regiões Sul, Sudeste e Centro-Oeste e região do MATOPIBA (Maranhão, Tocantins, Piauí e Bahia). Nos resultados de ambos modelos, o choque unitário no número de apólices produziu uma resposta negativa na área cultivada para os períodos subsequentes. Ou seja,

um número maior de apólices pode ter levado a uma redução da área cultivada. Mas se essa foi a única razão, então torna-se questionável o quanto o seguro rural é efetivo na manutenção do investimento traduzido na área cultivada.

Contudo, o mesmo choque apresenta uma resposta positiva na variável produtividade a partir do segundo período analisado. Sob a premissa de que mais concorrência pode levar, direta ou indiretamente, a uma maior produtividade e considerando que os seguros podem ajudar a manter ativos produtores que teriam saído do mercado sem o seguro, é possível inferir que o PSR contribui positivamente para o ambiente econômico da produção agrícola. Outro ponto que complementa a argumentação favorável à utilização de seguros agrícolas, é que um choque positivo no número de produtores e na importância segurada produzem respostas positivas na produtividade por no mínimo dois períodos, o que pode estar relacionado com o aspecto concorrencial abordado pelos autores.

A partir de modelos de séries temporais aplicados a dados obtidos da SUSEP, Adami e Ozaki (2012) estabelecem uma relação entre a propensão a contratação de seguros e o incentivo obtido por meio da subvenção do PSR. Os autores apontam que o programa é efetivo no fomento do mercado de seguros em que atua, e nesse sentido, a falta de recursos para o PSR pode prejudicar o funcionamento do sistema.

2.3 Elementos Atuariais

Nesta seção, serão definidos os elementos atuariais básicos que compõem o sistema do PSR, com foco especial na modalidade agrícola, que é o objeto de estudo deste trabalho.

No Brasil, a regulamentação do mercado de seguros é regida pelo Decreto-Lei nº73 Brasil (1966), no qual a Superintendência de Seguros Privados (SUSEP) e o Conselho Nacional de Seguros Privados são instituídos como órgãos oficiais que constituem o Sistema Nacional de Seguros Privados em conjunto com as sociedades autorizadas a operar em seguros privados, resseguradoras e corretores habilitados.

2.3.1 O Contrato de Seguro

A oficialização de um seguro é feita por meio de um contrato, no qual o segurador se obriga a garantir o interesse do segurado contra os riscos segurados, podendo ser anulado caso haja ato doloso das partes envolvidas, por exemplo, a adulteração ou omissão de provas, evidências ou outras circunstancias que influenciem no pagamento da indenização (BRASIL, 2003).

Há dois limites importantes na definição de um contrato de seguro: o Limite Máximo de Indenização (LMI) e o Limite Máximo de Garantia (LMG) das coberturas, os

quais são utilizados para a definição do objeto do seguro. Há quatro tipos de cobertura:

- **Custeio**: o qual pode ser acionado quando a produtividade obtida em uma lavoura não é suficiente para a cobrir os custos da área e cultura segurados;
- **Produtividade e Preço**: sendo o LMI baseado na produtividade garantida e no preço estabelecido no contrato de seguro, sendo este o preço de referência para indenização;
- **Faturamento ou receita**: neste caso, o produto conta com uma cobertura na qual o LMI é calculado com base no faturamento a ser obtido com a produção, sendo o preço dos contratos comercializados na B3 a referência para este cálculo, diferentemente da cobertura de Produtividade e Preço; e
- **Índices Climáticos**: Temperatura e Precipitação estão entre os principais índices utilizados para o cálculo do LMI desse tipo de cobertura, os quais são apurados e, caso verificadas as condições de sinistro, a indenização pode ser realizada dependendo também das demais condições contratuais.

Há duas formas de contratação desses seguros: Multirriscos e Riscos Nomeados. No primeiro, são cobertos diversos riscos climáticos em uma única cobertura; no segundo, os riscos cobertos são definidos no contrato com coberturas distintas. Entre as variáveis de interesse estão: *Produtividade Esperada, Nível de Cobertura, Riscos Excluídos, Coberturas Adicionais* e o *Valor da Saca*.

2.3.2 Mutualismo

Mutualismo é um conceito que se refere a uma forma de organização em que os membros de um grupo se unem com um objetivo comum, geralmente para obter benefícios mútuos. Isso geralmente é feito por meio da criação de uma organização, como uma cooperativa ou uma sociedade mútua.

No contexto do seguro, o mutualismo é a ideia de que os membros de uma seguradora se unem para compartilhar os riscos uns dos outros. Em uma seguradora mutualista, os membros pagam prêmios para a seguradora, que usa esse dinheiro para pagar sinistros e cobrir os custos operacionais. Se houver excedente de dinheiro no final do ano, ele é devolvido aos membros da seguradora na forma de dividendos ou descontos nos prêmios do próximo ano.

O mutualismo é uma forma de seguro baseada na solidariedade e na ajuda mútua. Ele se diferencia do modelo comercial tradicional, em que uma empresa de seguros tem o objetivo de maximizar o lucro para seus acionistas. As seguradoras mutualistas

têm como objetivo fornecer cobertura acessível e de alta qualidade para seus membros, enquanto equilibram a gestão dos riscos e os custos operacionais.

2.3.3 Cobertura

Cobertura, no contexto de seguros, é o termo usado para descrever a proteção financeira fornecida por uma apólice de seguro em caso de perda ou dano. Quando uma pessoa adquire uma apólice de seguro, ela está efetivamente transferindo o risco de uma perda financeira para a seguradora. Em troca, a seguradora se compromete a fornecer cobertura em caso de um evento coberto pela apólice. Essa cobertura pode incluir danos a propriedade, lesões corporais, responsabilidade civil, perda de renda e muitos outros tipos de riscos.

Dessa forma, a cobertura pode ser descrita em termos de seus limites e exclusões. O limite é o valor máximo que a seguradora pagará em caso de uma reivindicação, enquanto as exclusões são situações específicas que não estão cobertas pela apólice.

É importante que os segurados entendam claramente os termos de sua cobertura para garantir que estão adequadamente protegidos em caso de um evento inesperado. Além disso, é importante lembrar que diferentes apólices de seguro oferecem diferentes níveis de cobertura e que o prêmio do seguro pode ser afetado pelo nível de cobertura escolhido.

Ainda de acordo com a Lei Brasil (2003), cabe ao Comitê Gestor Interministerial do Seguro Rural definir as coberturas dos seguros participantes do PSR, inclusive a determinação de quais modalidades e culturas serão atendidas além de estabelecer os limites financeiros.

2.3.4 Sinistro, Vistorias e Indenização

Ao identificar perdas relevantes em sua lavoura, incorridos por eventos cobertos pelo seguro, o produtor deve comunicar imediatamente seu agente de seguros para que a área possa ser vistoriada por um perito designado pela seguradora.

Para garantir a cobertura do sinistro, o produtor deve cumprir suas obrigações conforme estabelecido na apólice do seguro, incluindo seguir as datas de plantio e recomendações do Programa de Zoneamento Agrícola de Risco Climático (ZARC). Assim, o ZARC permite ao produtor seguir os períodos com menor risco climático para implantação de sua lavoura, e portanto, para os seguros que cobrem essas áreas. Por esse motivo, as seguradoras requerem que os agricultores sigam rigorosamente essa metodologia, visando reduzir os riscos da atividade e obter menores valores de prêmio.

Há três tipos principais de vistoria durante a vigência da apólice: Prévia,

Monitoramento e Sinistro, senda a última dividida por critérios específicos, conforme ilustrado na Figura 3.

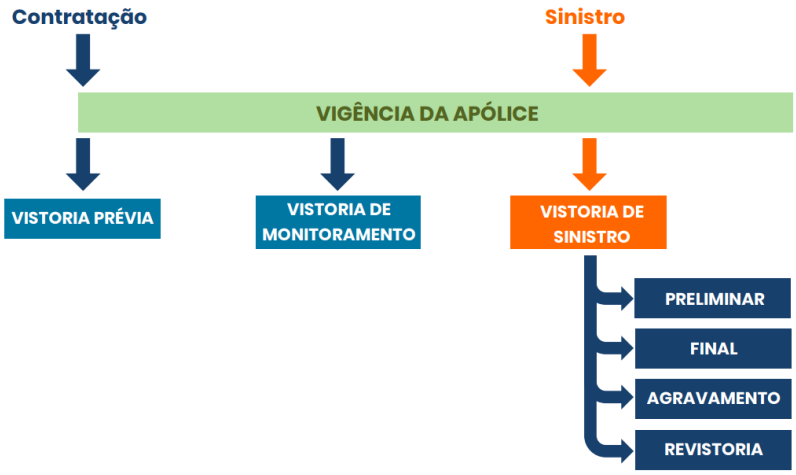


Figura 3 – Exemplo de Processo de Auditoria.

Fonte: MAPA (2020)

- **Prévia**: Essa vistoria permite verificar as condições do objeto segurado antes da efetivação do seguro. O objetivo é avaliar a situação inicial e determinar a aceitação do risco;
- **Monitoramento**: Realizada durante o período de vigência do seguro, essa vistoria tem como propósito acompanhar e monitorar as condições do objeto segurado ao longo do tempo. Ela pode ocorrer periodicamente para verificar possíveis mudanças ou eventos que possam afetar o risco segurado;
- **Preliminar**: Realizada no momento da ocorrência de um sinistro, essa vistoria tem como objetivo avaliar os danos iniciais e realizar uma análise preliminar dos prejuízos cobertos pelo seguro;
- **Final**: Realizada após a ocorrência de um sinistro, essa vistoria tem como objetivo verificar os danos finais sofridos pelo objeto segurado. Ela é essencial para determinar o valor da indenização a ser paga ao segurado;
- **Agravamento**: Realizada quando ocorre um evento que agrava o risco segurado, essa vistoria tem como objetivo avaliar as mudanças ocorridas no objeto segurado e analisar a continuidade do seguro ou a necessidade de ajustes nos termos do contrato; e
- **Revistoria**: Essa vistoria é realizada para revisar ou verificar novamente as condições do objeto segurado. Pode ocorrer para fins de renovação do seguro, avaliação de mudanças ocorridas ou atualização das informações cadastrais.

Além das vistorias, é importante que o produtor cumpra todas as obrigações estabelecidas na apólice do seguro, como permitir a inspeção dos bens segurados, comunicar circunstâncias que possam afetar o risco e apresentar as notas fiscais dos insumos utilizados na área segurada. Essas medidas visam garantir a correta avaliação e processamento do sinistro pelo programa de seguro rural no Brasil. O cumprimento das datas de plantio do ZARC também é uma obrigação dos produtores que contrataram apólices de seguro rural, pois isso garante menores riscos à atividade e menores valores de prêmio aos produtores.

2.3.5 A formalização da comunicação de sinistro

Os COPs ou Comunicados de Ocorrência de Perdas são protocolos que oficializam a comunicação de potenciais sinistros incorridas por eventos relacionados às coberturas contratadas.

2.4 Aspectos Psico-Sociais

A escolha feita pelo produtor de submeter uma proposta ou indenização pode ser influenciada por conhecidos efeitos da psicologia social e de outras áreas do conhecimento (BABCOCK, 2015), por exemplo, da aversão à perda(KAHNEMAN; KNETSCH; THALER, 1991) e do otimismo exagerado (DICKS; GARVEN; HILLIARD, 2018).

A fim de se proteger das perdas incorridas, o segurado pode, por exemplo, solicitar uma indenização mesmo que as causas não estejam de acordo com as coberturas do seguro. Nesse sentido, foi realizado um experimento que demonstra que indivíduos tendem a mentir para evitar resultados indesejados (GARBARINO; SLONIM; VILLEVAL, 2019). O segundo efeito é o do viés otimista e ele pode ser ilustrado em situações nas quais o produtor decide não contratar um seguro e tanto o evento coberto quanto as perdas incorridas acabam sendo confirmados (KAHNEMAN; LOVALLO, 1993). Outros fatores podem influenciar a decisão do produtor em efetivar seu contrato de seguro: um exemplo é a elasticidade preço-demanda (KRASHENINNIKOVA et al., 2019), mas estão além do escopo deste trabalho.

A variedade de produtos, tipos e níveis de cobertura pode impactar as decisões tomadas pelos potenciais segurados. Trata-se de um fenômeno chamado *choice overload*, que já foi explorado no campo de seguros de saúde por (ERICSON; SYDNOR, 2017).

Em seus experimentos, na área de Seguros de Saúde, Kairies-Schwarz et al. (2017) ressaltam a importância de utilizar a Teoria de Prospecto Cumulativo (*Cumulative Prospect Theory*) (TVERSKY; KAHNEMAN, 1992) para avaliar a qualidade de decisões. Em resultados empíricos, as preferências denominadas CPT, ou seja, aquelas que seguem *Cumulative Prospect Theory* (CPT) foram 91% das escolhas avaliadas na amostra, contra apenas 9% do tipo *Expected Utility Theory* (EUT).

Além disso, o sistema no qual ocorre a avaliação da indenização também está sujeito a outros problemas conhecidos que ocorrem no julgamento humano, como em sistemas ruidosos (KAHNEMAN et al., 2021). Nesse sentido, sinistros com condições relativamente parecidas entre si, podem ter resultados diferentes, tanto na decisão sobre a indenização ocorrer ou não, quanto na variância dos valores aprovados para indenizados deferidas.

Fraudes podem impactar o sistema de seguridade e é de particular interesse para a sociedade brasileira a exposição ao risco de fraude que recebem subvenção do governo, como o PSR. No âmbito das decisões mecanizadas, há uma variedade de modelos de *Machine Learning* que têm sido utilizados como estratégia para melhoria de decisão na Economia, em particular, na Prevenção contra Fraude. Algumas das áreas que podem ser destacadas: Cartão de Cédito, Telecomunicações e Seguros de Vida. Além das áreas de aplicação, os métodos mais relevantes são as Redes Neurais Artificiais, modelos baseados em Árvores de Decisão, Máquinas de Vetor de Suporte, *Naive Bayes* e KNN (SINAYOBYE; KIWANUKA; KYANDA, 2018).

Mais especificamente, em estudos sobre Redes Neurais Artificiais, há também métodos de Aprendizado Profundo por Reforço que são utilizados na área de Prevenção contra Fraude (MOSAVI et al., 2020), sendo de particular interesse para este trabalho as abordagens baseadas em *Q-Learning*, *Deep Q-Learning* e *Double Deep Q-Learning*, as quais tiveram sua efetividade em detecção de fraudes demonstrada com aplicação em dados de seguridade pública dos Estados Unidos, (CHOI; KIM; KIM, 2021) [4]. Outra abordagem de interesse, aplicada na área de Fraudes em operações de Cartão de Crédito, é a de Ambientes Adversários: o agente fraudador interage com o ambiente e pode ser ensinado a vencer o modelo anti-fraude atuante nesse ambiente (MEAD et al., 2018). Como fraudes estão em mudança o tempo todo, o Aprendizado Adaptativo poder ser usado para obter um auditor automático anti-fraude, o que pode ser uma vantagem, dadas as mudanças na dinâmica do ambiente e ganhos de eficiência de fraudadores (LU, 2010). Os algoritmos treinados para deteção de fraude tiveram a função de recompensa combinada com a distribuição teórica de Benford com três digitos [5] para obter a detecção de potenciais fraudes.

2.4.1 Inconsistências e Risco de Fraude

Após deferidas, algumas indenizações podem apresentar inconsistências ou divergências que podem comprometer o processo indenizatório. Caso seja verificado que ela foi reprovada em pelo menos uma das regras que compõe as políticas da seguradora e legislação em vigor, há indício de que esse potencial deferimento pode ter sido afetado

[4] Detalhes sobre as métricas que indicam essa efetividade são demonstrados. Os autores não estabelecem claramente a metodologia de amostragem dos dados de validação, por exemplo, validação cruzada. pelos autores, como o uso da métrica *F1 Score*.

[5] Detalhes sobre esse parâmetro presentes no estudo de referência

por uma fraude. Há particularidades que variam ao longo do tempo conforme políticas distintas são aplicadas para definir que uma indenização ou proposta tenha sido fraudada.

De acordo com o Código Civil brasileiro, o contrato de seguro e todas as suas obrigações podem ser anulados caso seja comprovada a fraude, por exemplo, por omissão dolosa. Ou seja, existem consequências diretas para o segurado que por ventura tenha cometido fraude para obter vantagens do sistema de seguridade.

A fundamentação teórica proposta nesta pesquisa foi desenvolvida com utilizando elementos da Economia, Ciências Atuárias e da Computação bem como Engenharia Agrícola. Nesse sentido, uma potencial redução da ineficiência causada por fraudes em um sistema de seguridade, no PSR em particular, é desejada como consequência da aplicação da metodologia proposta. Ao mesmo tempo, esta pesquisa expande a construção de conhecimento acerca do processo de decisão, o qual pode ser útil para a comunidade científica e para a sociedade civil.

2.5 Modelagem

Nesta sessão, são definidos os elementos e processos utilizados para implementação do ambiente de forma a capturar a dinâmica de construção das carteiras de indenização, na perspectiva da instituição seguradora que avalia as solicitações e executa as transações financeiras relativas ao processo de indenização. As políticas de indenização que são adotadas pelas seguradoras podem ser definidas como um conjunto de funções que, quando avaliadas, produzem o resultado da decisão para cada política vigente. No modelo proposto, o resultado é aprovação ou rejeição da solicitação de indenização. Para aplicação da lógica de auto-aprendizado proposta neste trabalho, as seguintes premissas precisam ser consideradas:

- Não mais de uma política, ou seja, um conjunto de regras, pode ser adotada para cada contrato na mesma decisão para:
  - Aprovação ou rejeição da proposta
  - Deferimento ou indeferimento da solicitação de indenização
- A mesma política que define fraude, avaliada por auditores especialistas diferentes, pode apresentar divergência, ou seja, a produção de variabilidade indesejada ou ruído.

Por razões frequentemente associadas à regulamentação e política comercial, seguradoras podem ajustar suas políticas, em diferentes *cohorts* [6] de forma que a cada

[6] Grupo com características em comum, como o mesmo período de aquisição do seguro

renovação da política vigente são feitas alterações nos parâmetros das funções que a compõem. Outra possibilidade é analisar os diferentes períodos de aquisição de contratos, nos quais o risco de sinistros futuros é estimado e precificado. A renovação contratual com precificação baseada em algoritmos de reforço foi porposta por (KRASHENINNIKOVA et al., 2019).

Além de um desafio técnico para estudos científicos nas Atuarias, a falta de dados que permitam inferir uma política, ou seja, se uma determinada indenização foi considerada oficialmente uma fraude traduz um aspecto prático da dinâmica do mercado de seguros: no momento da decisão sobre a indenização, a assimetria de informação inerente ao processo define a sua ocorrência *ex-post*.

A probabilidade de fraude como variável teórica é abordada a partir da função de Distribuição de Probabilidade de Benford, a qual é utilizada para calcular a função de recompensa. Com isso, fica estabelecida a relação do modelo probabilístico com a abordagem por Aprendizado de Máquina, especificamente, por Reforço.

Para determinar se um algoritmo treinado é eficaz na seleção de contratos a serem indenizados, é necessário verificar se as recompensas acumuladas aumentam ao longo dos episódios, o que indica uma melhor performance e, consequentemente, uma redução na probabilidade de fraude. As métricas tradicionalmente usadas serão abordadas na metodologia, mas a recompensa acumulada, métrica de referência nos estudos de Aprendizado por Reforço, será o principal guia para estabelecer a eficácia dessa abordagem.

2.5.1 Aprendizado por Reforço

No diagrama apresentado na Figura 1, há dois momentos de interesse para este trabalho, ambos diretamente ligados com a subvenção do PSR: a análise da proposta e a da indenização. No primeiro, ao prover a subvenção dos valores dos prêmios para a contratação do seguro, e portanto, um impulso na demanda por meio da diminuição relativa dos preços; no outro, pelo impacto que potenciais fraudes possam ter na precificação dos prêmios, o que pode ser mitigado por um processo de indenização mais eficiente na proteção dos valores dedicados ao programa.

Entre as diferentes estratégias para mitigar os riscos de fraude, existem os sistemas de inferências baseados em modelos estatísticos ou de Aprendizado de Máquina. Por exemplo, é possível utilizar uma classificação de probabilidade de fraudes associadas a uma Política Anti-Fraude, estabelecida *a priori*, e submeter casos de alta probabilidade para um analista humano.

A partir de uma revisão sistemática da literatura de Aprendizado de Máquina no para Detecção de Fraudes, Sinayobye, Kiwanuka e Kyanda (2018) identificaram as principais áreas, técnicas e vantagens de diferentes abordagens. Entre as áreas identificadas,

os seguros automotivos e de saúde respondem por quase 20%. Ou seja, existe uma relevância do tema de seguros quando são consideradas tarefas de detecção de fraudes. No âmbito técnico, as Redes Neurais Artificiais e modelos baseados em Árvores de Decisão aparecem como os algoritmos ou abordagens mais utilizados. Características como tolerância a variáveis interdependentes e velocidade para inferência da classificação foram apontadas como vantagens desses algoritmos. Ao mesmo tempo, apresentam desvantagens quanto aos requisitos de treinamento como o tamanho da amostra e a complexidade computacional (SINAYOBYE; KIWANUKA; KYANDA, 2018).

Outros métodos de Redes Neurais, como as Redes Profundas ou *Deep Learning*, podem ser utilizados em conjunto com a dinâmica do Aprendizado por Reforço para treinar modelos em diferentes tarefas no âmbito de fraudes, como a detecção e estimação do risco a eles associados. Na análise de redes sociais, por exemplo, foi possível implementar redes neurais de colisões indiretas para identificar ciclos de tráfego de veículos automotivos com baixa probabilidade de ocorrência. Com isso, é possível formular estratégias financeiras mais adequadas com a realidade do mercado de seguros automotivos. Além disso, o aprendizado por reforço já foi aplicado para detecção de fraudes na notificação de acidentes automotivos e na obtenção de uma técnica de codificação para dados de telemática. Esta última, chamada *auto-encoder*, foi implementada para obter perfis de riscos e estilos de direção, permitindo uma melhor adequação de produtos de seguros aos segmentos da demanda (MOSAVI et al., 2020).

Algumas empresas aplicam estratégias baseadas modelos de Aprendizado de Máquina para a detecção de fraude antes mesmo de incorrer em prejuízo financeiro contabilizado, como no mercado de cartões de crédito. As transações podem ser rejeitadas sem que seus respectivos valores monetários sejam operacionalmente comprometidos. Porém, a constante adaptação e criatividade dos fraudadores contra o sistema de seguridade exige uma dinâmica de atualização desses modelos que pode não ser atendida prática, sobretudo quanto a frequência de treinamento. Nesse contexto, Mead et al. (2018) construíram um agente fraudador utilizando simulações de Monte Carlo cujo objetivo é driblar as inferências obtidas com um modelo de regressão logística treinado para detecção de fraudes. Assim, o agente é recompensado positivamente ao conseguir realizar transações de baixo e alto valor sem serem rejeitadas conforme as recomendações obtidas a partir de um modelo de detecção de fraude. Os autores apontam que pequenas e consistentes mudanças no modelo de detecção de fraudes estão fortemente relacionadas com a capacidade de fraudadores aprenderem a driblar o sistema anti-fraude.

2.5.2 Elementos de Aprendizado por Reforço

Conforme ilustrado na Figura 4, no Estado presente S_t, o Agente toma ações que afetam o ambiente, passando para um novo Estado S_{t+1} e recebendo uma Recompensa R_t relativa à ação tomada.

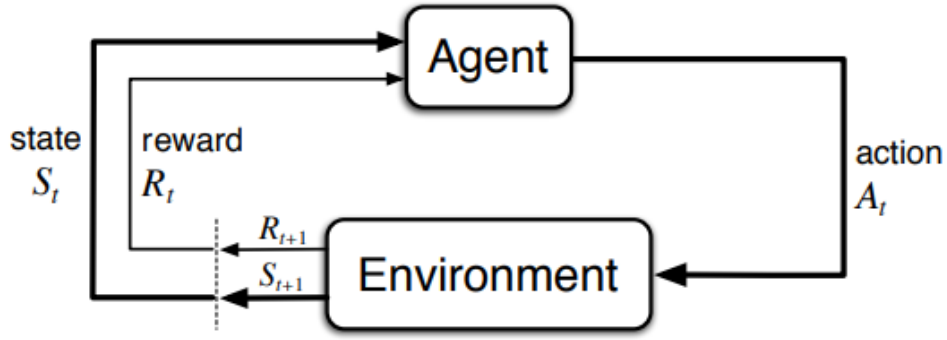


Figura 4 – FONTE: (SUTTON; BARTO, 2018)

Assim, o objetivo do Agente é maximizar as recompensas cumulativas sobre os diferentes Estados e para isso é preciso aprender a selecionar suas ações. O conjunto de ações de um agente é denominado de Política, sendo que ela pode ser determinística *a priori* ou estocástica, obtida ao longo do treinamento. A inferência dessa política pelo algoritmo é considerada uma forma de aprendizado denominada auto-aprendizado, no qual a recompensa é inferida sem a utilização de uma variável reposta ou target, mas com uma proxy ou aproximação que permita o aprendizado.

Neste contexto de aprendizado por reforço, cinco conceitos são essenciais: o agente, responsável pelas decisões e interações; o ambiente, que gera recompensas baseadas nas ações do agente; estados, que descrevem a situação atual do ambiente; ações, representando as escolhas do agente; e recompensas, que *fornecem* feedback às ações do agente, podendo ser imediatas ou atrasadas. Esses elementos formam a base para compreender como os agentes de aprendizado por reforço operam em uma ampla gama de aplicações.

- **Agente**: entidade que toma as decisões (ações) e interage com o ambiente;
- **Ambiente**: o sistema de estados que gerencia as recompensas de acordo com as interações com o Agente;
- **Estado**: Descrição do ambiente no tempo t, contendo os dados que traduzem o ambiente e sua situação atual;
- **Ação**: representação das escolhas realizadas pelo agente em resposta aos Estados do Ambiente. De acordo com o objetivo, as ações podem ser discretas ou contínuas; e

- **Recompensa ou reforço**: sinal emitido pelo ambiente em resposta a uma ação realizada pelo agente, podendo ser, por exemplo, imediato ou atrasado.

Já o caminho definido pelo conjunto de ações, estados e recompensas é chamado de trajetória, ou seja, a historia da interação do agente com o ambiente. Alguns ambientes possuem um Estado Terminal, ou seja, a sequência de estados possui um limite definido. Por exemplo, um ciclo de renovação de seguros pode ser considerado um episódio, pois possui um Estado Terminal. Dessa forma, a precificação de prêmios para renovação de seguros pode ser considerada uma tarefa episódica.

Krasheninnikova et al. (2019) basearam-se nessa abordagem para construir uma metodologia *model-free* de Aprendizado por Reforço para o problema de renovação dos prêmios de seguros com dados reais da seguradora do Banco Bilbao Vizcaya Argentaria. Ao final da vigência de um contrato, o segurado elegível tem a opção de renovar o seguro, ou seja, contratar uma nova apólice com as mesmas características [7] e obter a cobertura por mais um período ou ciclo. Porém, ao realizar a oferta de renovação, a precificação dos prêmios traz o desafio de equilibrar a retenção de clientes e as receitas da empresa.

Dessa forma, os autores propõem a utilização de variáveis no nível cliente (f_1^n), ou seja, considerando cada situação individual dos clientes, como a segmentação. Além disso, foram utilizadas variáveis no nível da empresa, denominadas pelos autores de *global features* (f_2^n) e que capturam a interação do agente com o ambiente por meio de métricas que traduzem os objetivos de retenção de clientes e crescimento de receita. Além disso, preços (*price*) e segmentação (*tier*) foram utilizados na composição dos estados do ambiente. Assim, cada estado do ambiente proposto é composto por uma tupla:

$$S_n = (f_1^n, f_2^n, price_n, tier_n)$$

Os autores realizam dois experimentos principais, sendo o primeiro com o objetivo de obter a maximização das receitas e, no segundo, o mesmo objetivo porém sujeito à restrições relativas à retenção dos clientes. Um ponto relevante que eles trazem e que é aproveitado neste trabalho é que abordar o problema de renovação de seguros como um Processo de Decisão de Markov parece adequado, o que foi demostrado pelos resultados comparando o fator de desconto γ, o qual determina a importância de recompensas futuras. Assim, foi demonstrado que o agente prefere sacrificar recompensas de curto-prazo com a maximização da receita para cada observação individualmente em favor de obter uma receita total de médio e longo prazo, portanto, mais adequadamente abordado como um problema de decisões em sequência.

De forma análoga, as solicitações de indenização de seguros agrícolas podem ser agrupadas em períodos de renovação da carteira, os quais são definidos pelas diferentes

[7] Existem varias especificidades da renovação dependendo da cobertura, legislação, entre outros fatores.

safras e culturas, políticas vigentes para o PSR, por exemplo. Além disso, a comunicação de sinistro pode acontecer em diferentes momentos durante o período da vigência da apólice, conforme as condições estabelecidas para o produto de seguros em questão. Ou seja, é possível estabelecer um ponto inicial e final de uma tarefa de detecção de fraudes episódica, na qual a composição de carteiras de indenização pode ser modelada como uma sequência de decisões a respeito de cada indenização solicitada ao longo do tempo.

Embora o agente proposto por Lu, Boritz e Covvey (2006) também mapeie as transações potencialmente fraudulentas e evidencie o problema como uma sequencia de decisões, sua proposta difere em diversos momentos da solução trazida pelos ambientes "adversariais"de Mead et al. (2018). Uma das diferenças é que o agente constrói um grafo de ações baseadas nas variáveis do problema, realizando a transação com o valor já registrado nos dados e não há um valor predeterminado de transação. Os experimentos realizados por Lu, Boritz e Covvey (2006) foram obtidos com dados de seguros de automotivos e de saúde, fornecidos por Ernest and Young e Manulife, respectivamente.

Contudo, a principal formulação trazida por Lu, Boritz e Covvey (2006) de interesse para este trabalho é a função de recompensa construída a partir da distribuição de probabilidade de Benford. Os detalhes desta distribuição são abordados na seção de construção do ambiente e, cabe observar que, ao utilizar essa distribuição, os autores conseguiram treinar o agente em uma tarefa de detecção de fraudes sem utilizar os dados de *Grount Truth* como variável resposta para as transações fraudulentas.

Dessa forma, as recompensas são dadas a partir da estimação da distribuição empírica de Benford para os três primeiros dígitos nos dados observados do valor de compra contra a distribuição teórica de Benford, conforme a equação (2.1).

$$Reward(i) = \frac{f_{1i}}{b_{1i}} + \frac{f_{2i}}{b_{2i}} + \frac{f_{3i}}{b_{3i}} \tag{2.1}$$

Cada termo da soma será mais próximo da distribuição de Benford quando o resultado da fração $\frac{f_{1i}}{b_{1d}} = 1$, em que d é o dígito sendo analisado. Quanto mais distante da distribuição teórica, mais distantes de 1 estarão os valores.

Assim, os autores conseguiram identificar a empresa de onde fraudes eram geradas utilizando a técnica de *Policy Evaluation* no ambiente que propuseram. Eles também argumentam que métodos tradicionais de detecção de outliers com a estatística de Benford, aplicada diretamente aos valores de transação, não teria permitido a identificação desta empresa.

Na discussão de resultados, é apontado que o agente conseguiu aprender esse padrão de fraude por ter ligado transações com fraudadores que submetiam as indenizações em lugar da empresa com as suas próprias localizações.

Ainda na área da saúde, com a mesma tarefa de detecção de solicitações de indenização fraudulentas e utilizando dados do *Medicare*, sistema de seguros de saúde dos Estados Unidos, Choi, Kim e Kim (2021) treinaram agentes DQN e DDQN [8] para classificar transações nas seguintes categorias: Normal, Normal Suspeita, Fraude, Fraude Suspeita. Com isso, as métricas obtidas nos experimentos apresentam um avanço em relação ao modelo concorrente (contemporâneo ao estudo). Por exemplo, houve um ganho de 3% na acurácia e 23% em precisão dos casos fraudulentos, além de melhorar a métrica *recall*, ou revocação, dos mesmos casos em 20% em comparação com o modelo concorrente.

Nesta seção, foram apresentados os elementos da literatura que dão suporte à construção do ambiente proposto neste trabalho, a saber: referencial teórico do Aprendizado por Reforço na detecção de fraudes, a seleção de indenizações como um problema de sequência de decisões, a utilização da distribuição de Benford para construir recompensas bem como o potencial do uso de agentes DQN e DDQN. Na seção seguinte, esses elementos serão formalizados e seus detalhes serão explorados para então implementar experimentos com o caso prático do PSR.

A fim de estabelecer o racional para as técnicas propostas nesta seção, é necessário compreender o potencial impacto de fraudes na dinâmica do mercado de seguros em um sistema como o PSR, isto é, como a ocorrência de fraudes impacta um sistema de seguridade.

Neste trabalho, é proposto um mercado de seguros teórico que foi desenvolvido para ser logicamente auto-contido, mantendo-se apenas os elementos fundamentais para estabelecer a existência do impacto de fraudes no mercado de seguros. Sendo assim, a proposta limita-se a demonstrar a influência de fraudes sobre os preços dos prêmios ao longo do tempo, tendo em vista a importância das renovações contratuais para esses mercados. Contudo, não é objetivo desta sessão exaurir a temática que deve ser aprofundada na vasta literatura de áreas como Economia, Ciênciais Atuariais e Estatística, entre outras.

Além disso, é apresentada a abordagem de auto-aprendizado por reforço, a qual permitirá que o modelo identifique padrões e anomalias nos dados sem depender de uma definição prévia específica de fraude. Essa abordagem possibilita que o sistema se adapte e aprenda com novos casos e comportamentos potencialmente fraudulentos, incluindo a dimensão temporal.

2.6 O impacto de fraudes

Nesta seção, será proposta uma construção lógica a partir dos elementos e relações definidos adaptados da literatura apresentada na introdução deste trabalho. Assim,

[8] Agentes de Deep Q-Networks e Double Deep Q-Networks, metodologias de Aprendizado por Reforço conforme descrito na Revisão de Literatura

as premissas nele propostas permitirão abstrair o sistema de inferências necessário para estabelecer a ***existência do impacto de fraude*** em um sistema de seguridade fictício. Não é objetivo, contudo, estabelecer o tamanho deste impacto, pois isso prevê uma gama de hipóteses além do escopo de interesse. Dessa forma, serão focos deste raciocínio: o contexto para o desenvolvimento do ambiente de Aprendizado por Reforço e como utilizar os modelos treinados para a prática de mitigação de riscos de fraude.

2.6.1 Um mercado de seguros fictício

Suponha-se um mercado de seguros de mutualismo com apenas uma seguradora, um tipo de seguro para um bem único; e um número n fixo de segurados que possuem apenas uma apólice, sendo que para este exercício é irrelevante a cobertura e detalhes específicos deste seguro. É necessário abstrair a taxa de sinistralidade, portanto que seja totalmente fixa para os períodos em consideração.

Seja PP_t a variável que representa os *Preços (nominais) dos Prêmios no tempo* t. Sejam O_t o *Orçamento* nominal estabelecido para custeio das indenizações deste seguro no tempo t. Seja SO_t o número total de *Sinistros do Orçamento no tempo* t para os quais o Orçamento O_t está definido. Seja S_t o número total *Sinistros* nesse mercado.A situação em que $S_t = SO_t$ ocorre quando o número de sinstros desse mercado SM_t é a mesma do previsto em orçamento, então esta é uma situação de equilíbrio.

Seja definido o *Valor Total do Mercado* como $VTM_t = PP_t * n$, ou seja, o preço nominal dos prêmios no tempo t multiplicado pelo número de segurados. Ao fazer o *Valor de Indenizações de Sinistros* VIS_t multiplicado pelo número total de Sinistros S_t no tempo t, chega-se no *Valor Total de Sinistros* VTS_t. Neste exercício, existe pelo menos um período de renovação dado por $t* = t + 1$, ou seja, um ciclo de renovação.

A pergunta de interesse para este exercício é: qual o valor dos prêmios no período de renovação? Caso não haja processo inflacionário algum neste mecardo imaginário, então o preço é dado por:

$$PP_{t+1} = PP_t \tag{2.2}$$

Ou seja, os preços são os mesmos para o período de renovação já que foi arbitrariamente desenhado para isso. Contudo, se $S_t > SO_t$, então existem indenizações pagas que estavam fora do orçamento. Para este exercício, sejam estas as indenizações definidas como fraudulentas. Por ora, é irrelevante saber a definição em si, basta considerar que foram indenizações pagas sem que estivessem orçamento deste mercado. Existem, portanto, SNO_t indenizações que não fazem parte do orçamento, dado por

$$SNO_t = S_t - SO_t$$

Sendo $VTS_t > O_t$ um resultado indesejado para a saúde deste mercado, pois essas indenizações trariam prejuizos, já que o orçamento é feito justamente para cobrir apenas os sinistros previstos [9].

Uma consequência desses prejuízos é o impacto no preço de renovação de seguros. Supondo que as fraudes incorridas tenham sido o único processo inflacionário, a expressão $VIS_t * SNO_t$ define o valor total de gastos fora do orçamento: $VTSNO_t$.

Caso os custos sejam absolutamente absorvidos pelo próprio mercado, no período de renovação, o novo orçamento O_{t+1} terá de incluí-los no cálculo dos prêmios. É possível, assim, precificar apenas o período imediatamente anterior, por exemplo, para compensar as perdas e acomodar os novos seguros, tendo em mente que o número de segurados desta hipótese é fixo.

$$PP_{t+1} = PP_t + \Delta PP_t \tag{2.3}$$

A partir de agora, o termo Δ, que representa a variação dos preços no tempo t, traz o ponto de partida da variáção dos preços para PP_{t+1} na medida em que este é tomado como uma proporção sobre os preços daquele, no período anterior, o que fica evidente se o domínio de $D(\Delta) \in \mathbf{R}$.

Isto é, os preços do período de renovação PP_{t+1} serão compostos pelos termos PP_t, o preço no periodo t imediatamente anetior a $t + 1$, somado de uma variação sobre PP_t. Essa variação é uma proporção do próprio preço, ou seja, $\Delta(PP_T) \in \mathbf{R}$ sendo que $PP_t * \Delta(PP_T) \in \mathbf{R}$. Assim, no escopo deste exercício, os preços futuros são os preços atuais somados a uma variação deles mesmos, podendo assumir valores negativos no caso de deflações.

Neste modelo, as fraudes são o único processo inflacionário atuante sobre os preços do período de renovação, especificado pelo termo ΔPP_t. Nesse sentido, atuar para previnir o pagamento de indenizações fraudulentas é o objetivo de tarefas de deteção de fraudes operacionalizadas *a priori*, ou seja, com reflexo prático e contábil na administração do orçamento de indenizações dos seguros.

Os agentes tomam diversas decisões ao longo da jornada de vida de um contrato de seguro que podem impactar o risco de fraude. Em particular, as seguradoras passam pela análise das solicitações de indenização, podendo indeferir um pedido que esteja em desacordo com as suas políticas, conforme explicado anteriormente. Nesse momento, a aplicação de técnicas de prevenção contra fraude é evidentemente relevante, já que ao evitar o pagamento da indenização no indeferimento da solicitação, a potencial ocorrência

[9] Em caso práticos, contudo, não é possível extrair a probabilidade, o que faz com que isso possa não ser verdade na vida real

de fraude pode ser investigada e, caso verificada, diminuir a exposição ao risco (financeiro) de fraude *a posteriori*.

A proposta deste trabalho é obter um ambiente computacional para treinar modelos que ajudem a construir carteiras de indenização livres de fraude e, de acordo com o racional apresentado, contribuir para a estabilidade e eficiência do mercado de seguros agrícolas.

2.7 Detectando o desconhecido

A constante reinvenção das formas e instrumentos aplicados para obter vantagens e burlar sistemas de seguros exige que a prevenção contra fraudes também esteja em busca de inovação dos mecanismos práticos de detecção. Mais especificamente para modelos de Aprendizado de Máquina, o fenômeno chamado *Drift* ocorre quando mudanças na distribuição dos dados impactam a performance de modelos, podendo ser causados por diversas situações como as alterações nos sistemas de seguridade.

No PSR, por exemplo, há diversas mudanças sistêmicas que podem acarretar em *Drifts* [10] como os diferentes orçamentos e limites de subvenção. Além disso, a dinâmica econômica, desastres naturais e as próprias fraudes podem trazer muitos desafios e efeitos de *Drifts*. Dessa forma, se os modelos aplicados não forem automaticamente atualizados com re-treinos, aprendizado ativo e *online*, por exemplo, suas métricas de performance podem ser impactadas diretamente e resultar na inviabilização do modelo como instrumento de detecção de fraudes e até prejuízos financeiros.

Salvo particularidades específicas de cada mercado de seguros, mudanças nas definições de fraude podem ocorrer antes, durante ou depois da ocorrência de fraude: a descoberta de novos fatos que levem à indenização a ser considerada fraudulenta pode também servir para avaliar o passado em *hindsight* [11] e novas definições de fraude para o futuro.

Faz-se necessário, portanto, obter técnicas que permitam realizar a detecção de casos potencialmente fraudulentos independentemente de uma definição de fraude. No entanto, como detectar se uma indenização será paga indevidamente sem necessariamente dar conta de todas as possíveis formas que uma fraude pode ser aplicada? Outra forma de colocar esta pergunta e que introduz o problema do ponto de vista da Ciência de Dados: como detectar casos potencialmente fraudulentos sem que os dados da variável resposta estejam disponíveis?

Assim, fica explicita a relação entre *Drifts*, a ausência de uma definição de

[10] Processo no qual mudanças nos dados inviabilizam o uso das inferências obtidas por modelos de Apendizado de Máquina

[11] A percepção de um evento após a sua ocorrência

fraude e a impossibilidade de obter variáveis-resposta que possam dar conta de todas as possíveis definições de fraude para um fenômeno. Neste trabalho, portanto, não é feita uma formalização da definição de fraude já que o interesse é poder detectar os casos de forma agnóstica a definições. Com isso, tome-se o risco de fraude como a variável teórica de interesse para este trabalho, uma vez que ela traduz a exposição do orçamento financeiro de um sistema de seguridade ao risco de realização de pagamentos de indenizações fraudulentas. Contudo, é necessário medir esse risco para permitir experimentação empírica, análise e comparação das diferentes abordagens candidatas a resolver o problema: a variável operacional deste trabalho é a probabilidade de fraude de carteiras de indenização, a qual é aproximada utilizando a distribuição teórica de Benford, como apresentado a seguir.

2.7.1 Análise de Dígitos

Faz-se necessário definir os elementos matemáticos e estatísticos que são utilizados nos próximos passos e associar esses objetos às técnicas de treinamento dos algoritmos. Para isso, são utilizadas as definições e a linguagem apresentadas no trabalho de Berger e Hill (2011). A definição a seguir é a referência para o que é tratado como ***primeiros dígitos*** ao longo deste trabalho, ou seja, os dígitos $j \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ que compõem um número real x, ordenando a identificação da esquerda para direita e iniciando a contagem a partir do primeiro dígito diferente de zero.

Definição 1 (Algarismos Significativos: Primeiros Dígitos)**.** *Para todo $x \neq 0 \in \mathbf{R}$, o primeiro digito decimal significativo de x, denominado $D_1(X)$, é o único inteiro $j \in \{1, 2, .., 9\}$, tal que $10^k j \leqslant |x| < 10^k(j+1)$ para algum $k \in \mathbf{Z}$ único.*

Da mesma forma, para cada $m \geqslant 2$, $m \in \mathbf{N}$, o m-ésimo digito decimal significante é definido, de forma indutiva como o único $j \in \{0, 1, 2, ..., 9\}$, tal que:

$$10^k\left(\sum_{i=1}^{m-1} D_i(x)10^{m-i} + 1\right) \leqslant |x| < 10^k\left(\sum_{i=1}^{m-1} D_i(x)10^{m-i} + j + 1\right) \tag{2.4}$$

para algum $k \in \mathbf{Z}$ necessariamente único.

Com essa definição, pode-se obter os diferentes dígitos de um número real. Por exemplo, fazendo $x = \pi = 3,14$, o primeiro digito é dado por $D_1(\pi) = 3$ e de forma análoga, tem-se $D_2(\pi) = 1$, $D_3(\pi) = 4$. Outro exemplo seria um valor monetário de indenização, amostrado dos dados deste trabalho, $x = 16.642,56$, sendo $D_1(16642,56) = 1$, $D_2(16642,56) = 6$, $D_3(16642,56) = 6$, ou seja, o primeiro digito desse valor é 1, o segundo é 6 e o terceiro é 6.

2.7.2 Newcomb-Benford

O matemático e astrônomo Simon Newcomb publicou no *American Journal of Mathematics* o seu achado a respeito da frequência de dígitos ao estudar tabelas logarítmicas e observações sobre o estado de conservação de livros, inferindo a frequência nas quais as primeiras páginas ficam mais gastas do que as últimas. Em outro trabalho chamado A Lei dos Números Anômalos, Frank Benford apresentou a análise da distribuição dos dígitos de 20 conjuntos de dados de diversas áreas (MILLER et al., 2015).

Com eles, Benford demonstrou a ocorrência do Fenômeno do Primeiro Dígito ou Lei de Benford: $Prob(D_1 = d_1) = \log_{10}(1 + d_1^{-1})$ para todo $d_1 = 1, 2, ..., 9$. Ou seja, a probabilidade do primeiro digito observado nesses dados é uma distribuição logarítmica. Contudo, uma versão prática mais interessante para este trabalho é apresentada a seguir:.

Definição 2 (Lei Forte de *Benford* para os Primeiros Digitos)**.** *Seja $d \in \mathbf{Z}[1, 9]$ o número inteiro (também denominado digito) sendo avaliado, ou seja, aquele para o qual busca-se saber a distribuição de probabilidade. Seja $s \in \mathbf{R}$, a lei forte de Benford conforme definida por Miller et al. (2015) segue: Diz-se que um conjunto de dados satisfaz a Lei Forte de Benford se a probabilidade do significando ser observado no intervalo real $[1, s)$ é $\log_{10} s$*

Assim, em vez de analisar se o primeiro digito é um determinado valor, estende-se a lei para analisar a probabilidade do significando estar entre $[d, d + 1)$, obtendo-se

$$\log_{10}(d + 1) - \log_{10}(d) = log_{10}\frac{d + 1}{d} \tag{2.5}$$

O exemplo a seguir explicita a distribuição do primeiro digito para alguns casos. É notável como a distribuição varia conforme os dígitos crescem no conjunto de inteiros $1, 2, ..., 9$, sem contar o zero (à esquerda), uma vez que não faz diferença para existência nem unicidade dos números reais.

Exemplo 1 (Análise do Primeiro Dígito)**.** *Considerando apenas o primeiro digito e substituindo para diferentes j, tem-se a distribuição apontada na Tabela 2:*

Assim, é possível obter as probabilidades associadas aos dígitos de acordo com a distribuição de Benford. A Tabela 2 possui três colunas que descrevem esse procedimento, representando os dígitos de 1 a 9, fornecendo as probabilidades correspondentes arrendondadas para quatro casas decimais, sendo que as probabilidades são expressas em logaritmos na base 10.

Assim, ela é uma ferramenta fundamental para o trabalho e aplicação da distribuição de Benford, permitindo a compreensão das frequências esperadas para cada dígito em diferentes contextos de análise estatística e detecção de anomalias. A formulação

Tabela 2 – Distribuição Teórica de Benford

j	$\text{Prob}(D_1 = j)$	$\text{Prob}(D_1 = j)$	$\text{Prob}(D_1 = j)$
1	$\log_{10}\left(1+\frac{1}{1}\right)$	$\log_{10} 2$	0.3010
2	$\log_{10}\left(1+\frac{1}{2}\right)$	$\log_{10}\frac{3}{2}$	0.1760
3	$\log_{10}\left(1+\frac{1}{3}\right)$	$\log_{10}\frac{4}{3}$	0.1249
4	$\log_{10}\left(1+\frac{1}{4}\right)$	$\log_{10}\frac{5}{4}$	0.0969
5	$\log_{10}\left(1+\frac{1}{5}\right)$	$\log_{10}\frac{6}{5}$	0.0792
6	$\log_{10}\left(1+\frac{1}{6}\right)$	$\log_{10}\frac{7}{6}$	0.0669
7	$\log_{10}\left(1+\frac{1}{7}\right)$	$\log_{10}\frac{8}{7}$	0.0579
8	$\log_{10}\left(1+\frac{1}{8}\right)$	$\log_{10}\frac{9}{8}$	0.0513
9	$\log_{10}\left(1+\frac{1}{9}\right)$	$\log_{10}\frac{10}{9}$	0.4575

Fonte: Produzido pelo autor.

que permite a extensão da distribuição de Benford para acomodar todos os dígitos é apresentada a seguir, na Equação 2.6:

$$\sum_{k=10^{n-2}}^{10^{n-2}-1} \log_{10}(1+1/(10k+d)) \tag{2.6}$$

Exemplo 2. *O exemplo a seguir demonstra como usar o numero de digitos n e o digito d de interesse para avaliar a distribuição de Benford.*

1. *Para $n = 3$ e $d = 1$, a soma é:*

$$\sum_{k=10^{3-2}}^{10^{3-1}-1} \log_{10}(1+1/(10k+1))$$

2. *Para $n = 4$ e $d = 2$, a soma é:*

$$\sum_{k=10^{4-2}}^{10^{4-1}-1} \log_{10}(1+1/(10k+2))$$

3. Para $n = 5$ e $d = 3$, a soma é:

$$\sum_{k=10^{5-2}}^{10^{5-1}-1} \log_{10}(1 + 1/(10k + 3))$$

Para $n = 3$ e $d = 1$, a fórmula está somando os logaritmos na base 10 de $1 + 1/(10k + 1)$ para todos os k de $10^{3-2} = 10$ a $10^{3-1} - 1 = 99$. Isso significa que estamos considerando todos os números de dois dígitos e calculando a probabilidade de que o dígito 1 apareça em cada posição.

Para $n = 4$ e $d = 2$, a fórmula está somando os logaritmos na base 10 de $1 + 1/(10k + 2)$ para todos os k de $10^{4-2} = 100$ a $10^{4-1} - 1 = 999$. Isso significa que estamos considerando todos os números de três dígitos e calculando a probabilidade de que o dígito 2 apareça em cada posição.

Para $n = 5$ e $d = 3$, a fórmula está somando os logaritmos na base 10 de $1 + 1/(10k + 3)$ para todos os k de $10^{5-2} = 1000$ a $10^{5-1} - 1 = 9999$. Isso significa que estamos considerando todos os números de quatro dígitos e calculando a probabilidade de que o dígito 3 apareça em cada posição. Ao explorar a conexão entre o problema de fraudes e a distribuição de Benford, é importante considerar os seguintes elementos:

Análise dos n primeiros dígitos: A distribuição de Benford estabelece que os dígitos menores (1 a 9) devem ocorrer com maior frequência em números autênticos. Portanto, ao investigar fraudes, pode-se comparar a distribuição observada desses dígitos com a distribuição esperada de Benford. Desvios significativos podem indicar possíveis irregularidades.

A compreensão da lógica subjacente da distribuição de Benford é fundamental para a construção da lógica de aproximação dessa distribuição com o uso do Aprendizado de Máquina. Ao considerar esses elementos em conjunto, é possível estabelecer uma abordagem mais robusta para conectar o problema de fraudes à distribuição de Benford, permitindo a identificação de potenciais anomalias e aprimorando os esforços de detecção de fraudes.

2.7.3 Métricas de avaliação

O Aprendizado por Reforço surgiu como um paradigma poderoso para uma variedade de aplicações, desde jogos até robótica. Tradicionalmente, modelos de aprendizado de máquina são avaliados usando métricas como Precisão e Acurácia. No entanto, essas métricas, embora eficazes para tarefas de aprendizado supervisionado, podem não capturar os desafios e requisitos nuances do AR. No contexto de AR, a complexidade e dinâmica dos ambientes, juntamente com o ciclo contínuo de *feedback* de ações e resultados, exigem

métricas especializadas que ofereçam uma visão mais holística do desempenho de um algoritmo.

As métricas de Aprendizado por Reforço (POOLE; MACKWORTH, 2010) permitem avaliar o desempenho e o progresso do agente, abrangendo aspectos como recompensa, duração dos episódios, taxa de exploração e sucesso.Na Tabela 3 estão as métricas selecionadas para este trabalho:

Tabela 3 – Métricas de Aprendizado por Reforço

Métrica	Fórmula
Recompensa média por episódio	$\frac{1}{N}\sum_{i=1}^{N} R_i$
Recompensa acumulada por episódio	$\sum_{t=1}^{T} r_t$
Recompensa máxima por episódio	$\max_t r_t$
Recompensa mínima por episódio	$\min_t r_t$
Número médio de passos por episódio	$\frac{1}{N}\sum_{i=1}^{N} T_i$
Número total de passos por episódio	$\sum_{i=1}^{N} T_i$
Taxa de Falsos Positivos	$\frac{\text{Falsos Positivos}}{\text{Falsos Positivos + Verdadeiros Negativos}}$
Taxa de Falsos Negativos	$\frac{\text{Falsos Negativos}}{\text{Falsos Negativos + Verdadeiros Positivos}}$

Fonte: Elaborado pelo autor.

A **Recompensa média por episódio** é relevante porque fornece uma medida da média de ganhos ou perdas obtidos pelo agente ao longo dos episódios de treinamento. Já a **Recompensa acumulada por episódio** por episódio é útil para entender como as recompensas se acumulam ao longo do tempo, indicando o retorno acumulado das ações tomadas pelo agente.

Além disso, as **Recompensa máxima e mínima por episódio** são cruciais para identificar os melhores e piores desempenhos alcançados pelo agente durante o treinamento. Isso permite entender a variabilidade das recompensas obtidas e possibilita investigar o motivo por trás dos resultados extremos. Outra métrica importante é o

Número médio de passos por episódio, que representa o número médio de ações tomadas pelo agente para completar uma tarefa. Isso ajuda a entender o nível de esforço ou ação necessária para atingir um resultado desejado.

Além disso, o **Número total de passos por episódio** é relevante para analisar a quantidade total de ações tomadas pelo agente ao longo do treinamento. Essa métrica pode ajudar a entender a eficiência do modelo em relação ao número total de ações necessárias para concluir as tarefas de seleção de carteiras de indenização.

Por fim, **Falsos Negativos:** indenizações rejeitadas pelo agente quando na verdade levariam a distribuição da carteira de indenização para mais próximo da distribuição teórica de Benford. Eles ocorrem quando o agente de aprendizado por reforço rejeita indenizações que, na realidade, contribuiriam para aproximar a distribuição da carteira de indenizações da distribuição teórica de Benford. Isso significa que o agente pode erroneamente descartar indenizações legítimas que seriam compatíveis com a distribuição desejada. Esses falsos negativos resultam em uma perda de oportunidade de melhorar a conformidade da carteira com a distribuição de Benford. Portanto, é importante considerar o equilíbrio entre a detecção de fraudes e a inclusão de indenizações genuínas para maximizar a eficácia do aprendizado por reforço na construção da carteira de indenizações.

Falsos Positivos: indenizações aceitas pelo agente, mas que na verdade aumentam a distância da carteira de indenização aprovada da distribuição de Teórica de Benford. Falsos Positivos ocorrem quando o agente de aprendizado por reforço aceita indenizações que, na realidade, aumentam a distância entre a carteira de indenizações aprovada e a distribuição teórica de Benford. Isso significa que o agente pode erroneamente incluir indenizações que não são compatíveis com a distribuição desejada. Esses falsos positivos comprometem a conformidade da carteira com a distribuição de Benford, resultando em uma distorção indesejada. Portanto, é crucial estabelecer um equilíbrio entre a detecção de indenizações legítimas e a rejeição das indenizações que se afastam da distribuição teórica, a fim de garantir a qualidade e a integridade da carteira de indenizações construída pelo agente de aprendizado por reforço.

A questão de preferir entre falsos positivos ou falsos negativos depende do contexto específico e das consequências associadas a cada tipo de erro. Ambos têm implicações diferentes e possam ter impactos significativos no resultado final.

Por um lado, ter falsos positivos significa aceitar indenizações que não estão em conformidade com a distribuição teórica de Benford. Isso pode resultar em distorções e afetar a integridade da carteira de indenizações. Se a conformidade com a distribuição de Benford for um requisito crucial, é preferível minimizar os falsos positivos para manter a precisão e a confiabilidade da carteira. No entanto, ao reduzir os falsos positivos, existe o risco de rejeitar indenizações genuínas que estariam de acordo com a distribuição desejada, o que pode levar a perdas de oportunidade e impactar negativamente o desempenho geral

da carteira.

Por outro lado, ter falsos negativos significa rejeitar indenizações que, na verdade, seriam compatíveis com a distribuição teórica de Benford. Isso pode levar a uma perda de oportunidade de melhorar a conformidade da carteira e potencialmente diminuir a eficácia do aprendizado por reforço. Se a detecção de fraudes for uma prioridade absoluta, minimizar os falsos negativos pode ser mais importante para evitar a inclusão de indenizações fraudulentas. No entanto, ao reduzir os falsos negativos, existe o risco de incluir indenizações que se afastam da distribuição de Benford, afetando a qualidade e a integridade da carteira.

Em resumo, a decisão entre preferir falsos positivos ou falsos negativos depende das prioridades e objetivos específicos do sistema de seguros. É essencial equilibrar a detecção de fraudes com a inclusão de indenizações genuínas para maximizar a eficácia e a precisão da construção da carteira de indenizações. Uma abordagem adequada envolve avaliar cuidadosamente os riscos e impactos de cada tipo de erro e encontrar um equilíbrio que atenda às necessidades do sistema de seguros e dos clientes envolvidos. Assim, essas métricas fornecem informações valiosas sobre o desempenho do agente e sua capacidade de selecionar carteiras de indenização com base em critérios específicos.

Recompensas cumulativas servem como a métrica primária de avaliação em muitas tarefas de AR. Elas representam os benefícios agregados que um agente recebe ao longo do tempo, orientando o agente a tomar decisões que maximizem os retornos a longo prazo. No entanto, focar exclusivamente em recompensas acumulativas pode ser enganador. Elas podem não capturar a variabilidade no desempenho, riscos potenciais ou a estabilidade de um algoritmo, especialmente quando implantado em ambientes diferentes ou em constante mudança.(CHAN et al., 2019)

Portanto, para garantir a robustez e confiabilidade de algoritmos de AR, pesquisadores e profissionais estão voltando sua atenção para métricas abrangentes de avaliação. Essas métricas visam fornecer uma compreensão mais profunda do comportamento de um algoritmo, enfatizando aspectos como variabilidade, risco e consistência durante as fases de treinamento e pós-aprendizado.

3 Metodologia

Neste capítulo, aborda-se a fundamentação teórica juntamente com a implementação computacional do Ambiente de Aprendizado *InsurEnv*. Com isso, são detalhados os elementos essenciais do Aprendizado por Reforço, os quais são cruciais para a elaboração de carteiras de indenização com um risco reduzido de fraude. Especificamente, o foco recai sobre o mercado de seguros agrícolas, que emerge como a principal motivação desta pesquisa.

Paralelamente, a dinâmica de aprendizado e as recompensas são formuladas tendo como base a distribuição de **Benford**. Esta abordagem tem o propósito de habilitar a detecção de possíveis indenizações fraudulentas, alinhando-se à premissa de manter carteiras de indenização saudáveis.

Na sequência, descrevem-se os experimentos realizados no ambiente proposto. A primeira estratégia adotada envolve uma PO (Política Operacional) aleatória, visando avaliar a estabilidade e consistência do ambiente em diferentes cenários. Por outro lado, a segunda estratégia centra-se no uso do Agente DQN, com o intuito de analisar a eficácia do aprendizado no ambiente e proporcionar uma avaliação mais detalhada dos resultados alcançados.

Esses experimentos, por sua vez, são fundamentais não apenas para validar a viabilidade e eficácia do ambiente proposto, mas também para aprofundar o entendimento sobre a detecção de fraudes em carteiras de indenização.

3.0.1 A distribuição de Benford das Carteiras de Indenização

A inferência que materializa a variável operacional deste trabalho é que ***carteiras de indenização livres de suspeita de fraude seguem a distribuição de Benford***. Isso sugere que a distribuição dos valores nas carteiras de indenização genuínas segue uma tendência estatística específica, característica da distribuição de Benford. Essa coerência com o padrão de Benford pode ser um indicativo de integridade e ausência de manipulação nos dados das carteiras de indenização. Portanto, a análise da distribuição de Benford pode ser uma ferramenta útil na detecção e prevenção de fraudes nesse contexto, proporcionando maior confiança e segurança nos processos de seguros agrícolas.

Essa forma de inferir a saúde dessas carteiras quanto ao risco de fraude permite o treinamento de algoritmos para detecção de potenciais fraudes sem que seja necessário

ter os dados da variável resposta, mas além disso, sem uma definição arbitrária de fraude.

A capacidade de treinar algoritmos para detectar fraudes sem depender de dados da variável resposta ou de definições arbitrárias de fraude é revolucionária. Uma ferramenta que se destaca nesse contexto é a distribuição de Benford, definida como:

Definição 3. *A distribuição de Benford (BERGER; HILL, 2011)* $\mathbf{B}$ *é a única medida de Probabilidade sobre* $(\mathbf{R}^+, S)$*. Para todo* $m \in \mathbf{N}$*, todo* $d_1 \in \{1, 2, ..., 9\}$ *e todo* $d_i \in \{1, 2, ..., 9\}$ *com* $j \geqslant 2$*:*

$$\mathbf{B}(D = d_i) = \log(1 + (\sum_{i=1}^{m} 10^{m-i} d_i))^{-}1 \tag{3.1}$$

para $i \in \{1, 2, ..., m\}$

Quando a probabilidade observada para cada dígito da carteira de indenizações é igual à probabilidade teórica dada pela distribuição de Benford, isso indica que a carteira segue exatamente o padrão esperado e é considerada ideal nesse contexto. Assim, pode-se definir uma carteira ideal C_p, da seguinte forma:

Definição 4 (Carteira Ideal)**.** *Seja* C_p *um subconjunto de* $\mathbf{R}^+$ *que representa os valores das indenizações de uma carteira de pagamentos, na qual cada* p *é o índice do contrato em sendo analisado. Se a probabilidade observada* $\mathbf{B_o}(D = d_i)$ *for igual a Probabilidade Teórica* $\mathbf{B}(D = d_i)$ *de Benford, diz-se que esta é uma Carteira Ideal* CI *e seus* p *contratos são os contratos ideais para um determinado número* m *de dígitos, os quais compõem a carteira de indenizações.*

Essa igualdade entre as probabilidades observadas e teóricas é um critério possível para determinar se uma carteira de indenizações se enquadra na categoria de carteira ideal, conforme estabelecido no trabalho.

O trabalho de Lu, Boritz e Covvey (2006) foi a principal inspiração para chegar na proposta trazida neste trabalho, tanto na utilização da distribuição de Benford quanto na dinâmica do aprendizado ao utilizar iteração de estados para cada transação analisada. Contudo, cabe verificar que esta metodologia de auto-aprendizado é limitada, pois há necessidade de avaliar a distribuição observada para variáveis coletadas desde que sigam a distribuição de Benford[1].

Além da distribuição, é fundamental definir como as sequências de decisões a serem tomadas pelo Agente vão permiti-lo construir as carteiras de acordo com a distribuição de Benford, uma vez que há 2^n possíveis escolhas para tanto.

[1] Na ausência de dados monetários, áreas e volumes, por exemplo, outra forma de auto-aprendizado pode ser necessária para substituir a da distribuição de Benford. Mesmo com algumas dessas variáveis, outras condições como a disponibilidade de dados em termos de número amostral e condições sistêmicas podem trazer desafios para utilização desta metodologia.

Essa sequência de decisões é de particular interesse, pois motiva a utilização de inteligência artificial e aprendizado de máquina, na medida em que permitem estimar o risco de fraude de forma probabilística, ao mesmo tempo em que permitem obter variáveis dependentes que podem ser usadas para interpretação dos modelos estimados.

No universo da detecção de fraudes, a sequência de decisões desempenha um papel crucial. A motivação para incorporar inteligência artificial e aprendizado de máquina nesse processo é evidente, visto que essas tecnologias possibilitam a estimativa probabilística do risco de fraude e fornecem variáveis interpretativas para os modelos. Uma abordagem matemática que complementa essa análise é a sequência de Benford, cuja definição é apresentada a seguir:

Definição 5 (Sequências de Benford)**.** *Uma sequência é chamada de Benford se o limite da frequencia relativa dos digitos analisados for uma distribuição de Benford conforme a equação abaixo (BERGER; HILL, 2011):*

$$\lim_{N\to\infty}\left(\frac{\#(1 \leqslant n \leqslant N \ tal\ que\ D_k(x_n) = d_i)}{N}\right) = \log\left(1 + \left(\sum_{i=1}^{m} 10^{m-i} d_i\right)^{-1}\right) \tag{3.2}$$

- $D_i(x_n)$ é o i-ésimo dígito do n-ésimo número na sequência.
- A função $\#(1 \leqslant n \leqslant N$ tal que $D_i(x_n) = d_j)$ conta o número de vezes que o dígito d_j aparece na i-ésima posição entre os primeiros N números da sequência.
- O termo $\sum_{j=1}^{m} 10^{m-j} d_j$ é a representação decimal dos primeiros m dígitos.

3.0.2 Processo de decisão

Conforme explorado entre as referências utilizadas, o problema de seleção de indenizações para composição de carteiras ideais pode ser modelado como uma sequência de decisões. Assim, uma carteira será formada a partir da aprovação ou seleção de um contrato para pagamento. Nesse sentido, toma-se como premissa para o ambiente proposto que, o valor de indenização solicitado é o mesmo do reportado nos dados oficialmente divulgados pelo MAPA e utilizados neste trabalho.

Seja p uma proposta e C_p o conjunto de solicitações de indenização. Seja C_p^* o conjunto de indenizações aprovadas. Podemos definir a função de ação $A(p)$ da seguinte maneira:

$$A(p) = \begin{cases} 1 & \text{se } p \in C_p^*, \\ 0 & \text{caso contrário.} \end{cases} \tag{3.3}$$

Esta função atribui o valor 1 a $A(p)$ se p estiver em C_p^* (ou seja, a proposta é aprovada), e 0 caso contrário (ou seja, a proposta é rejeitada).

A tomada de decisões em qualquer sistema complexo envolve uma série de ações sequenciais que, juntas, definem a estratégia operacional do sistema. Para entender e representar matematicamente essas sequências de ações, introduz-se o conceito de Política Operacional. Esta política encapsula todas as possíveis sequências de decisões que podem ser tomadas. A definição matemática de uma Política Operacional é detalhada abaixo:

Definição 6 (Política Operacional)**.** *Seja s um número natural que representa a contagem das ações tomadas em uma sequência de decisões A_s:*

$$A_s = \{a_1, a_2, \ldots, a_t\}$$

Uma PO (Política Operacional) é o conjunto formado pelas diferentes sequências A_s possíveis:

$$PO := \{A_1, A_2, \ldots, A_s\}$$

onde cada A_s é uma sequência de decisões.

Assim, se t for um número real fixo maior que 1, a sequência é finita, representando o estado terminal.

Ou seja, o máximo da cardinalidade desse conjunto pode ser obtida por:

$$\max card(PO) = 2^t$$

uma vez que as únicas ações possíveis são aprovar ou rejeitar a solicitação de indenização. Como o conjunto $C_p^* \in \mathbf{R}$, então sempre é possível obter a distância do máximo para todos os elementos do conjunto PO, o que é diretamente verificável: suponha que existe um elemento a_t* para o qual a função $\max_d |P_{\text{observada}}(D_i = d) - P_{\text{teórica}}(D_i = d)|$ não é definida. Isso implicaria que existe uma decisão a_t* para a qual a probabilidade observada $P_{\text{observada}}(D_i = d)$ ou a probabilidade teórica $P_{\text{teórica}}(D_i = d)$ não é bem definida.

Porém, as funções de distribuição de probabilidade definidas nesta seção são bem definidas na sequência de números reais positivos. Isso contradiz a suposição inicial é de que todas as probabilidades observadas e teóricas são bem definidas para todas as decisões possíveis. Portanto, existe um elemento a_t^* para o qual a função $\max_d |P_{\text{observada}}(D_i = d) - P_{\text{teórica}}(D_i = d)|$ não é definida e deve ser falsa.

Isso completa a demonstração de que a função:

$$\max_d |P_{\text{observada}}(D_i = d) - P_{\text{teórica}}(D_i = d)|$$

é bem definida para todos os pontos de A_s.

3.1 A tarefa de seleção de Carteiras Ideiais

Em Reinforcement Learning (RL) ou Aprendizado por Reforço, uma trajetória, também conhecida como episódio ou sequência de ações, é uma sequência de estados, ações e recompensas que um Agente experimenta ao interagir com um ambiente.

Assim, ela se inicia com o Agente observando o estado inicial do ambiente. Com base nesse estado, o Agente seleciona uma ação para executar. Em seguida, o Agente executa essa ação no ambiente, resultando em uma transição para um novo estado e na obtenção de uma recompensa. Esse processo se repete até que uma condição de término seja alcançada, como atingir um estado objetivo ou exceder um número máximo de etapas.

Seja uma trajetória representada por w, ela consiste em uma sequência ordenada de elementos $(s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, \ldots, s_{T-1}, a_{T-1}, r_T, s_T, a_T, r_{T+1})$, onde: - s_t representa o estado no tempo t, para $t = 0, 1, 2, \ldots, T$, - a_t representa a ação tomada no tempo t, para $t = 0, 1, 2, \ldots, T-1$, - r_{t+1} representa a recompensa obtida após tomar a ação a_t no estado s_t, para $t = 0, 1, 2, \ldots, T-1$, - T é o tempo final da trajetória.

Para a construção da carteira ideal, considerando o conjunto de todas as possíveis trajetórias T_w, a trajetória ótima (ou ótimas) é aquela que seleciona todos os w que minimizam a distância entre a distribuição empírica de Benford dos elementos do conjunto C_p^*, entre todas as possíveis tratetórias.

Entretanto, o conjunto T_w não é conhecido previamente, e a natureza estocástica do problema requer que o Agente explore diferentes opções e aja de forma a explorar regiões do espaço de indenizações que ainda não foram amostradas. Isso é importante para garantir que a distribuição das indenizações seja representativa e se aproxime da distribuição de Benford. À medida que o Agente explora e coleta mais amostras, ele pode atualizar sua política para se concentrar nas regiões que se aproximam da distribuição de Benford.

Uma trajetória é essencial para o processo de aprendizado em RL, pois fornece ao Agente as interações necessárias para aprender a tomar ações que maximizem a recompensa acumulada ao longo do tempo. O objetivo do Agente é aprender uma política ótima, ou seja, uma política que maximize a recompensa esperada ao longo de várias trajetórias. Assim, o objetivo do Agente nesse ambiente pode ser formulado como a seguir, na expressão 3.4 :

$$\text{minimize} \sum_{w \in T_w} \text{dist}(\mathbf{B}_{\text{teórica}}, \mathbf{B}_{\text{empírica}}(C_p^*)) \tag{3.4}$$

Sendo $p = p(w)$, ou seja, cada proposta p selecionada durante a trajetória w, para as quais $A(p) = 1$, por definição.

Em cada etapa da trajetória, o Agente recebe informações sobre o estado atual do ambiente, toma uma decisão com base em sua política de ação e recebe uma recompensa do ambiente. Essas informações são usadas para atualizar a estimativa de valor ou política do Agente e melhorar seu desempenho no aprendizado.

Esses argumentos fundamentam a utilização do Aprendizado por Reforço como uma abordagem adequada para esse contexto, proporcionando uma seleção mais precisa e consistente das indenizações. A aproximação de funções no Aprendizado de Máquina é o foco deste trabalho: uma técnica utilizada para estimar ou aproximar a função de valor em problemas complexos ou de grande escala. A função de valor representa a utilidade esperada de estar em um determinado estado e seguir uma determinada política de ação.

$$V(s;\theta) \approx \hat{V}(s;\theta) \tag{3.5}$$

Em vez de representar a função de valor de forma tabular, que seria inviável em problemas com grandes espaços de estado, essa técnica permite estimar funções usando uma função aproximadora parametrizada. Essa função aproximadora pode ser uma rede neural, uma função linear, uma árvore de decisão ou qualquer outra forma de modelagem que seja capaz de generalizar a partir de exemplos.

A ideia principal por trás da técnica *Value Function Approximation* é aprender uma representação compacta e geral da função de valor que permita ao Agente generalizar para estados não visitados durante o treinamento. Isso é especialmente útil em problemas complexos em que o espaço de estado é muito grande e seria impraticável ou impossível armazenar e atualizar uma tabela completa de valores para cada estado.

Ao utilizar a aproximação da função de valor, o Agente pode estimar a função de valor para qualquer estado, mesmo que não tenha sido visitado anteriormente, com base nas informações disponíveis. Essa aproximação permite ao Agente tomar decisões informadas, escolhendo ações que levem a estados de maior valor esperado, com o objetivo de maximizar a recompensa acumulada ao longo do tempo.

A escolha específica da função aproximadora e dos métodos de otimização depende do problema em questão e das características do ambiente e do Agente. O Aprendizado por Reforço oferece diversas abordagens para a *Value Function Approximation* como o uso de redes neurais profundas no Agente DQN (Deep Q-Network), que é um exemplo popular de aplicação dessa técnica.

Em resumo, *Value Function Approximation* é uma técnica que permite estimar ou aproximar a função de valor em problemas de Aprendizado por Reforço de forma mais

eficiente e escalável, possibilitando ao Agente generalizar o conhecimento aprendido e tomar decisões informadas em ambientes complexos.

3.2 O Ambiente *InsurEnv*

Nesse ambiente, o Agente possui duas opções: aprovar ou reprovar uma solicitação de indenização. Ao aprovar, o valor da indenização é deduzido do saldo da conta corrente do Agente. O espaço de estados é composto por variáveis históricas e variáveis de estado.

As variáveis históricas são calculadas a partir das datas de propostas, levando em consideração o histórico de cada segurado em relação às variáveis independentes extraídas dos dados originais. Essas variáveis fornecem informações sobre o histórico do segurado, como o número de propostas anteriores realizadas.

As variáveis de estado incluem o saldo da conta corrente do Agente, representado pelo capital de trabalho disponível (WC), e o desperdício de capital (WoC), que quantifica recursos financeiros utilizados de forma ineficiente ou em indenizações indevidas. Essas variáveis são cruciais para monitorar a saúde financeira do Agente e identificar oportunidades de melhoria na alocação de recursos.

As variáveis de estado são atualizadas a cada passo com base em uma função de recompensa que considera a diferença entre as probabilidades observadas e teóricas. O Agente busca maximizar essa diferença, buscando a aprovação eficiente de indenizações legítimas e minimizando o desperdício de capital.

Dessa forma, o Agente utiliza as variáveis de estado e a função de recompensa para ajustar sua estratégia e alcançar resultados desejados, como a otimização do orçamento e a maximização da eficiência no processamento de indenizações.

3.2.1 O Agente

O Agente representa o "avaliador"da solicitação e pode aprovar a indenização solicitada, ou seja, deferir o pedido de pagamento da indenização, subtraindo de sua conta corrente, o montante necessário. Alternativamente, o Agente pode reprovar a solicitação sem que o valor seja deduzido do saldo ou orçamento da carteira.

O objetivo do Agente nesse ambiente é obter a maior quantidade acumulada de recompensas possíveis (rewards) por meio da construção de diferentes carteiras de indenização. O Agente busca criar carteiras que se aproximem o máximo possível da distribuição de Benford, uma distribuição estatística que segue um padrão específico. A estratégia vencedora é aquela que consegue construir carteiras que se aproximam da distribuição de Benford e, ao mesmo tempo, alcançam a maior recompensa possível. Isso

implica em encontrar um equilíbrio entre seguir o padrão estatístico desejado e maximizar a quantidade de recompensas obtidas.

3.2.2 Estados, Passos e Episódios

O espaço de estados neste trabalho é composto por um conjunto de variáveis históricas (Vh) e variáveis de estado (Ve). As variáveis históricas foram extraídas dos contratos disponíveis na base de dados e submetidas a um conjunto de funções codificadas computacionalmente para obter o *dataset* final de treinamento. Essas variáveis históricas fornecem informações valiosas sobre o histórico dos contratos, permitindo que o Agente de Aprendizado por Reforço descubra padrões passados ao tomar suas decisões. As variáveis de estado, por sua vez, representam o estado atual do ambiente e do Agente, fornecendo informações atualizadas que influenciam a tomada de decisões. A combinação das variáveis históricas e de estado cria um espaço de estados abrangente e rico, que o Agente utiliza para descobrir e compreender os padrões presentes nos dados. Esse espaço de estados permite ao Agente tomar decisões mais informadas e adaptativas com base no contexto específico do ambiente e nos dados históricos disponíveis.

3.2.3 Estado: Variáveis Históricas

A partir das datas de proposta, foi realizado um processo de Engenharia de Variáveis, calculando a evolução histórica de cada segurado com respeito às variáveis independentes extraídas dos dados originais apresentadas no Apêndice A deste trabalho.

Cabe ressaltar que é esperado que haja mais de um registro por segurado, uma vez que o mercado de seguros possui o processo inerente de renovação e que permite obter variáveis que trazem informações do histórico desse segurado, por exemplo, a partir do número de propostas anteriormente realizadas por segurado, por data de proposta. O processo de engenharia de dados aplicado foi o cálculo da soma, o máximo e o mínimo de das variáveis numéricas independentes até a data variável, data da proposta, chamada *dt_proposta*, realizando a transformação e agregação de dados, conforme descrito abaixo.

Filtragem: O primeiro passo é filtrar o conjunto de dados para incluir apenas os registros até a data da proposta. Isso pode ser feito usando uma condição que verifica se a data de cada registro é menor ou igual à Data da Proposta (*dt_proposta*). O propósito dessa amostragem é separar uma porção de dados temporalmente extraídos para processar a validação retroativamente. No contexto deste trabalho, a validação de dados novos (não utilizados durante o treinamento) possui uma abordagem específica sendo recomendada para orientar a utilização do ambiente com dados do próprio PSR.

Agrupamento e Agregação: Após a filtragem dos dados, os dados seriam agrupados pelo ID da proposta e pela Data da Proposta (*dt_proposta*). Para cada ID

(*id_proposta*) da proposta, seria calculado a soma, o máximo e o mínimo das seguintes variáveis até a Data da Proposta (*dt_proposta*):

- **Valor do Prêmio Líquido (*vl_premio_liquido*):** Valor nominal em R$ pago à Seguradora equivalente à contratação da respectiva apólice;
- **Área Total (*nr_area_total*):** Área relativa à proposta ou contrato em questão;
- **Nível de Cobertura (*niveldecobertura*):** Grau de proteção fornecido pela apólice em relação aos riscos segurados;
- **Valor da Indenização (*valor_indenizacao*):** Valor da indenização realizado para a dada apólice;
- **Valor da Subvenção Federal (*vl_subvencao_federal*):** Valor nominal em R$ da subvenção dedicada à apólice;
- **Percentual de Taxa (*pe_taxa*):** Taxa percentual relacionada a algum aspecto do contrato;
- **Produtividade Segurada (*nr_produtividade_segurada*):** Número que representa a produtividade segurada pela apólice; e
- **Limite de Garantia (*vl_limite_garantia*):** Valor máximo garantido pela apólice.

Para a variável "Valor da Indenização"(*valor_indenizacao*), foi utilizada a codificação *One-Hot* para representar os dígitos da parte inteira do número real que representa o valor de indenização. Essa codificação cria variáveis binárias que indicam a presença ou ausência de cada dígito na parte inteira, capturando informações detalhadas sobre os dígitos presentes nos valores de indenização das observações em questão.

Essa codificação tem o objetivo de fornecer informações aos algoritmos de treinamento sobre os dígitos e seu impacto nas medidas de WoC e WC da carteira de indenização durante o processo de treinamento. Isso permite que o Agente aprenda como os diferentes dígitos influenciam a construção da carteira e tome decisões mais informadas com base nessas informações.

A variável evento preponderante também passou por uma codificação semelhante. Ela foi transformada em um conjunto de variáveis binárias, onde cada variável representa a presença ou ausência de um evento específico que é considerado preponderante. Essa codificação permite que o Agente leve em consideração a ocorrência desses eventos significativos durante o processo de tomada de decisão.

Essas codificações *One-Hot* enriquecem o espaço de estados do ambiente, fornecendo informações adicionais sobre os valores de indenização e os eventos preponderantes.

Essas variáveis codificadas ajudam o Agente a capturar e explorar padrões mais detalhados e específicos nos dados, melhorando sua capacidade de aprender e tomar decisões mais precisas.

Processos de Decisão de Markov [2] podem ser usados com ou sem restrição para modelar o ambiente de avaliação de seguros, pois assume-se que ações tomadas no presente afetam o futuro além de utilizar o valor esperado da recompensa por cada ação para o aprendizado (MEAD et al., 2018).

3.2.4 Estado: Variáveis do Episódio

A cada passo dentro do ambiente, o Agente tomará uma decisão que influenciará a atualização de quatro variáveis de estado importantes. A primeira delas é o Percentual de Orçamento Gasto, que reflete a porcentagem do orçamento disponível que foi utilizado até o momento. Essa variável permite ao Agente monitorar e controlar os gastos em relação aos recursos disponíveis.

A segunda variável é o Percentual de Indenizações Aprovadas, que indica a proporção de indenizações que foram aprovadas em relação ao total de indenizações solicitadas. Essa métrica é fundamental para avaliar o desempenho do Agente na seleção e aprovação de indenizações, fornecendo *insights* sobre a eficiência e a qualidade das decisões tomadas.

A terceira variável de estado, denominada WC (*With Current*), desempenha um papel fundamental no ambiente de Aprendizado por Reforço, refletindo o estado atual das médias das diferenças normalizadas entre a distribuição de Benford teórica e a empírica. Atualizada a cada passo do Agente, essa variável representa a média considerando o caso atual (p). Ao levar em conta a variável WC, o Agente é capaz de avaliar o desempenho em relação à distribuição observada e teórica, o que permite tomar decisões informadas para otimizar suas ações. Essa informação auxilia o Agente a aprender como influenciar a distribuição de eventos a fim de minimizar a média das diferenças normalizadas, buscando alcançar resultados superiores no ambiente de Aprendizado por Reforço. Ao trabalhar para minimizar a média das diferenças normalizadas, o Agente procura se aproximar o máximo possível da distribuição teórica desejada, aprimorando seu desempenho e maximizando sua habilidade de influenciar os resultados do ambiente.

With current considera o caso atual (p), Without Current não considera. A variável WoC (*Without Current*) representa o estado atual das médias das diferenças normalizadas entre a distribuição de Benford teórica e a distribuição empírica, sem levar em consideração o caso específico p sob avaliação no estado atual. Ao levar em conta a

[2] Um processo pode ser abordado ou adaptado para atender à propriedade de Markov como as PMDPs: *Partially Observable Markov Decision Processes* (KRASHENINNIKOVA et al., 2019)

variável WoC, o Agente é capaz de avaliar o desempenho global da distribuição empírica e tomar medidas para melhorar sua aderência à distribuição teórica.

Essas quatro variáveis de estado desempenham um papel crucial no acompanhamento e no gerenciamento do desempenho do Agente no ambiente de Aprendizado por Reforço. Por meio de sua atualização contínua a cada ação tomada, o Agente pode ajustar sua estratégia e buscar a maximização dos resultados desejados, como a otimização do orçamento, a aprovação eficiente de indenizações legítimas e a minimização do desperdício de capital.

As variáveis de estado do ambiente são atualizadas a cada passo com base em uma função de recompensa que considera a diferença entre as probabilidades observadas e teóricas. Assim, "WC"e "WoC"representam dois cenários diferentes para calcular a diferença absoluta máxima entre as probabilidades observadas e teóricas para um determinado dígito $D_i = d$. Nas Definições 7 e 8 a seguir, é esclarecido como esses elementos serão calculados:

Definição 7 (Distância de Benford com o Caso Atual)**.** *A Distância de Benford com o Caso Atual (WC) representa a distância da distribuição de Benford considerando o caso atual* (p)*. Ela é calculada utilizando a seguinte fórmula:*

$$\frac{1}{3}\sum_{i=1}^{3}\frac{\max_d |P_{observada}(D_i = d) - P_{teórica}(D_i = d)|}{||P_{observada}(D_i = d) - P_{teórica}(D_i = d)||}$$

Aqui, o numerador calcula a diferença absoluta máxima entre as probabilidades observadas e teóricas para um dígito específico $D_i = d$, enquanto o denominador calcula a distância euclidiana entre as probabilidades observadas e teóricas. O WC é utilizado como parte da função de recompensa para avaliar o desempenho do Agente e incentivar ações que aumentem essa distância quando o caso atual (p) é incluído.

Definição 8 (Distância de Benford sem o Caso Atual)**.** *A Distância de Benford sem o Caso Atual (WoC) representa a distância da distribuição de Benford considerando a exclusão do caso atual* (p)*. Ela é calculada utilizando a seguinte fórmula:*

$$\frac{1}{3}\sum_{i=1}^{3}\frac{\max_d |P_{observada}(D_i = d) - P_{teórica}(D_i = d)|}{||P_{observada}(D_i = d) - P_{teórica}(D_i = d)||}$$

Nessa fórmula, o numerador calcula a diferença absoluta máxima entre as probabilidades observadas e teóricas para um dígito específico $D_i = d$, enquanto o denominador calcula a distância euclidiana entre as probabilidades observadas e teóricas.

WC inclui o caso atual (p) sob avaliação no cálculo, enquanto *WoC* exclui o caso atual (p). A função de recompensa compara esses dois cenários e fornece uma recompensa baseada na proporção entre a diferença absoluta máxima com e sem o caso atual (p).

Se a diferença absoluta máxima for maior quando o caso atual (p) é incluído ($WC > WoC$), a recompensa é positiva e proporcional à razão $WC(s)/WoC$. Se for menor ($WC < WoC$), a recompensa é negativa e proporcional à mesma razão.

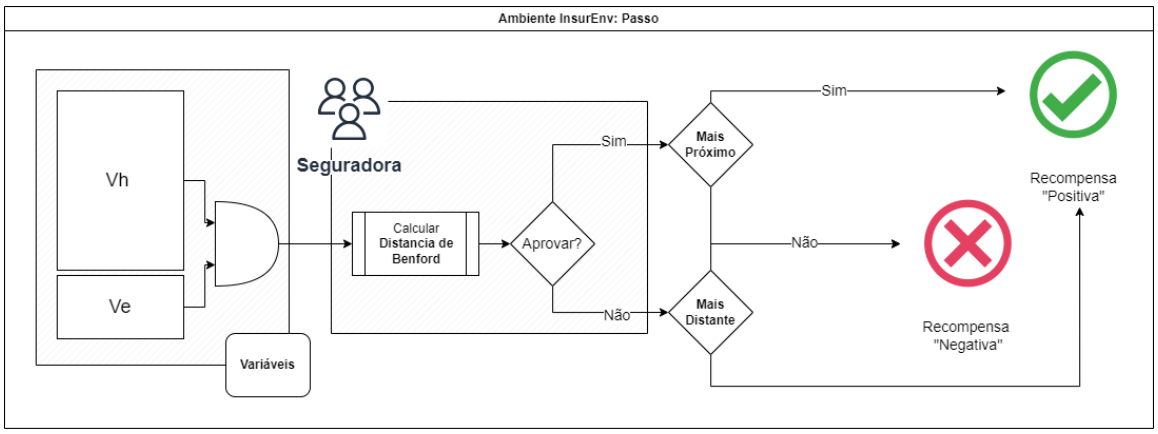
Se a diferença absoluta máxima for maior quando o caso atual (p) é excluído ($WoC > WC$), a recompensa é positiva e proporcional à razão $WoC/WC(s)$. Se for menor ($WoC < WC$), a recompensa é negativa e proporcional à mesma razão.

Essa função de recompensa parece incentivar um Agente a tomar ações que aumentem a diferença absoluta máxima entre as probabilidades observadas e teóricas quando o caso atual (p) é incluído, e diminuam essa diferença quando o caso atual (p) é excluído. Isso pode ser útil em um cenário onde se deseja que o Agente aprenda a influenciar a distribuição de eventos de uma maneira que maximize essa diferença.

3.2.5 Passos, casos e valor

Existe uma relação injetiva entre o conjunto de passos e casos. Dar um passo significa tomar uma ação (decisão) a respeito do caso relativo a este passo, conforme ilustrado na Figura 5. Cada passo corresponde a tomar uma ação ou decisão em relação a um caso específico. Isso significa que há uma correspondência unívoca entre os passos que o Agente realiza e os casos que são avaliados.

Figura 5 – Ambiente InsurEnv



Fonte: Elaborado pelo autor.

Essa relação bijetiva entre passos e casos permite que o Agente tome decisões individuais para cada caso em uma sequência definida de passos, portanto, traduzindo a sequência de decisões explorada na teoria apresentada neste trabalho. Cada passo avança o Agente para o próximo caso a ser avaliado, e a decisão tomada nesse passo afeta diretamente aquele caso em particular.

Dessa forma, o Agente progride por meio dos passos, tomando decisões em relação a cada caso individualmente, até que todos os casos tenham sido analisados e o episódio seja concluído. A relação bijetiva garante que cada caso seja tratado de forma única e que todas as decisões sejam aplicadas corretamente a cada caso correspondente.

Para a amostra de dados neste trabalho, o valor da indenização foi selecionado como variável de Benford a qual será utilizada para calcular as recompensas. Assim, valores de indenização potencialmente manipulados para driblar o sistema de seguridade podem ser identificados pelo algoritmo.

3.2.6 Parâmetros do Ambiente

Os parâmetros apresentados a seguir constituem atributos particulares [3] do ambiente que podem influenciar diretamente as decisões do Agente. Além disso, desempenham um papel crucial na definição da dinâmica de Aprendizado por Reforço, ou seja, um modelo da dinâmica do mercado de seguros para a tarefa de construção de sequencias de carteiras de indenização nas quais a detecção de possíveis irregularidades no processo de indenização é evidenciada.

- **Dados de Treinamento e Teste**: Refere-se aos conjuntos de dados utilizados para treinar e testar o modelo de Aprendizado por Reforço. Contudo, em abordagens de AR, a validação do algoritmo não é feita de forma tradicional como nos casos de modelos supervisionados, por exemplo;
- **Índice**: Variável que indica o nível de agregação dos dados. Esse índice pode ser utilizado para segmentar os dados em diferentes grupos ou categorias, permitindo uma análise mais detalhada e específica dentro do escopo;
- **Variável de Benford**: É um dos parâmetros da função de recompensa utilizada no modelo de Aprendizado por Reforço. Essa variável define quais dígitos serão avaliados em relação à distribuição de Benford pela função de recompensa a cada passo do processo de tomada de decisão;
- **Variáveis Históricas**: Compõem o espaço de estados (s) do modelo de Aprendizado por Reforço. Essas variáveis são baseadas em informações históricas relevantes e são utilizadas para representar o estado atual do ambiente e guiar as decisões do Agente;e
- **Orçamento**: Representa o valor inicial disponível para realizar o pagamento das indenizações. Esse orçamento é uma restrição que deve ser considerada durante o processo de seleção das indenizações, garantindo que as decisões tomadas estejam dentro dos limites orçamentários estabelecidos.

[3] Podem variar por país, região, mas são particulares à dinâmica de seguros.

A variável orçamento ou, *budget* como na implementação computacional, traduz uma importante premissa a respeito da dinâmica de avaliação das indenizações: o próprio PSR possui um valor orçamentário definido *a priori* pelo Congresso Nacional. Além disso, a utilização desse limitador permite o treinamento de Agentes que buscam maximizar as recompensas obtidas. A carteira do Agente guarda as informações de sua conta corrente e permite subtrair os valores destinados ao pagamento da indenização.

3.2.7 Episódio e Estado Terminal

Cada episódio termina caso uma de duas condições seja verificada: se o orçamento mínimo necessário para o pagamento da p-ésima indenização aprovada for insuficiente, ou se todos os casos a serem analisados tiverem sido concluídos.

A primeira condição indica que o episódio chega ao fim se o orçamento disponível não for suficiente para cobrir o pagamento da *p-ésima* indenização aprovada. Isso implica que o Agente não tem recursos financeiros adequados para continuar aprovando indenizações sem comprometer sua capacidade de pagamento. A segunda condição indica que o episódio termina quando todos os casos a serem analisados foram concluídos. Nesse ponto, o Agente considera que não há mais casos para avaliar e tomar decisões.

Ambas as condições marcam o fim de um episódio, ou Estado Final, no ambiente de Aprendizado por Reforço, e o Agente pode iniciar um novo episódio a partir desse ponto, com um orçamento e conjunto de casos atualizados, assim, o Estado Inicial possui 100% do orçamento intacto. A partir dela é possível obter o identificador de indenizações aprovadas pelo algoritmo e controlar a saída de recursos bem como o saldo disponível.

3.2.8 Função de Recompensa

No processo de avaliação de propostas e solicitações de indenização, há uma sequência de decisões a serem tomadas e essa caraterística enseja o uso de Aprendizado por Reforco. Além disso, há uma diferença temporal entre a aprovação de uma proposta no tempo t_0 e avaliação da solicitação de indenização que pode ocorrer em $t_1 | t_1 > t_0$, o que pode ser tratado como recompensa tardia para o treinamento do algoritmo.

Uma função genérica de recompensa em um problema de Aprendizado por Reforço, onde s representa o estado atual e a a ação tomada, pode ser expressa como na Equação 3.6:

$$R(s, a) = E[r_{t+1} | s_t = s, a_t = a] \tag{3.6}$$

Aqui, r_{t+1} é a recompensa recebida após a ação a ser tomada no estado s, e $E[r_{t+1}|s_t = s, a_t = a]$ é a expectativa dessa recompensa. Esta é uma definição genérica e pode ser adaptada para se adequar a qualquer problema específico de Aprendizado por Reforço.

Como uma ação define um novo estado para o ambiente em que o algoritmo está sendo treinado, foram desenhados experimentos considerando ações para estados imediatamente posteriores à ação, já que se trata de um processo estocástico e está sendo modelado com recompensas imediatas neste ambiente.

Neste contexto, WC (*With Current*) e WoC (*Without Current*) representam duas médias diferentes das diferenças normalizadas entre as probabilidades observadas e teóricas para um conjunto de eventos D_i, onde i corresponde aos *IDs* de propostas que foram aprovadas pelo algoritmo. $WC(p)$ inclui o caso atual sob avaliação (se aprovado) no cálculo, além dos casos já presentes na carteira atual $C_p^*(s)$. Por outro lado, WoC exclui o caso atual (p) do cálculo, considerando apenas os casos na carteira atual, os casos correntemente aprovados para pagamento.

Conforme definido em (3.7), $WC = WC(p)$ e $WoC = WoC(p)$ são as médias das diferenças normalizadas com e sem o caso atual (p) (se aprovado), respectivamente.

$$\text{Rewards}(WC, WoC, a, p) = \begin{cases} -\dfrac{WoC}{WC}, & \text{se } a_p = 0 \text{ e } WC \leqslant WoC, \\ \dfrac{WoC}{WC}, & \text{se } a_p = 1 \text{ e } WC \leqslant WoC, \\ \dfrac{WC}{WoC}, & \text{se } a_p = 0 \text{ e } WoC < WC, \\ -\dfrac{WC}{WoC}, & \text{se } a_p = 1 \text{ e } WoC < WC. \end{cases} \tag{3.7}$$

Essa função de recompensa incentiva o Agente a tomar ações que diminuam a média das diferenças normalizadas entre as probabilidades observadas e teóricas quando o caso atual (p) é incluído (se aprovado) nas indenizações pertencentes à carteira em construção C_p^*. Essa abordagem é útil em um cenário onde o objetivo é que o Agente aprenda a influenciar a distribuição de eventos de forma a minimizar essa média das diferenças normalizadas. Ao maximizar as recompensas acumuladas na contabilização geral dos episódios, o Agente busca otimizar seu desempenho ao longo do tempo, adotando estratégias que visem a construção da carteira ideal.

3.3 Configuração dos Experimentos

No podcast "Os Caminhos do Seguro Rural" (Sistema FAEP, 2023), Pedro Loyola, representante do Ministério da Agricultura no Governo Federal em 2022, compartilha informações relevantes sobre a atuação das organizações sociais Ocepar e Faep no Paraná. Essas entidades desempenharam um papel significativo ao promover e divulgar os seguros agrícolas apoiados pelo Programa de Subvenção ao Prêmio de Seguro Rural (PSR). Por meio de ações e estratégias dedicadas, eles buscaram conscientizar os agricultores sobre os benefícios dos seguros como ferramenta de gestão de riscos e sua eficácia no planejamento da safra.

Com o intuito de avaliar empiricamente a metodologia proposta, o escopo definido é o de ***indenizações da cultura de soja no estado do Paraná, no período de 2006 a 2022***, já que a ampla adoção e estratificação sistêmica desse grupo proporcionam uma quantidade significativa de dados. Esses experimentos permitem a análise e validação da abordagem proposta, utilizando dados reais e considerando as especificidades do mercado de seguros agrícolas no estado do Paraná, com potencial para generalização de maneira a abranger um conjunto maior de casos de Prevenção contra Fraude.

3.3.1 Política Aleatória

A fim de estabelecer a validade do ambiente para o Aprendizado por Reforço, foram realizados experimentos utilizando uma Política Aleatória. Nesta política, o Agente toma decisões de aprovar ou rejeitar a indenização solicitada a partir de uma distribuição aleatória com 50% de chance para cada opção.

Seja $\mathbf{O}$ o conjunto de todas as observações e p o número de seleções feitas pelo Agente. Para cada seleção k em $1, 2, \ldots, p$, o Agente escolhe aleatoriamente uma observação de $\mathbf{O}$ e toma uma decisão. Denotamos por C_k o número de decisões de aprovação feitas pelo Agente até a seleção k. A Distância de Benford (WC) na seleção k é então dada por

$$\text{BenfordDistance}_k = \frac{C_k}{k}.$$

Suponha que a variável de fato seja *Benford*, é possível verificar aplicando o Teorema do Limite Central, à medida que k se aproxima do infinito, BenfordDistance_k se aproxima de uma distribuição normal com média igual à Distância de Benford (WC) na população, denotada por BenfordDistance_p. Isso é expresso como

$$\lim_{k\to\infty} \delta\text{BenfordDistance(k, p)} = \text{BenfordDistance}_k - \text{BenfordDistance}_p = 0,$$

onde δ(BenfordDistance(k, p)) = BenfordDistance$_p$ − BenfordDistance$_k$ é a diferença entre a Distância de Benford (WC) na população p e a Distância de Benford (WC) na seleção k. Além disso, a Distância de Benford (WC) para cada seleção k deve se aproximar da Distância de Benford (WC) para a população à medida que k se aproxima do infinito. Isso é expresso como

$$\lim_{k \to \infty} \delta(\text{BenfordDistance(k, p)}) = 0,$$

onde δ(BenfordDistance) é a diferença entre a Distância de Benford (WC) para a população e a Distância de Benford (WC) para a seleção (MILLER et al., 2015).

Os experimentos realizados utilizando uma Política Aleatória servem como uma linha de base, ou "baseline", para a comparação do desempenho de um algoritmo de Aprendizado por Reforço. A premissa é que, se o algoritmo de Aprendizado por Reforço estiver aprendendo de maneira eficaz, ele deve ser capaz de fazer escolhas que resultem em uma Distância de Benford (WC) mais próxima da distribuição teórica de Benford do que a obtida pela Política Aleatória.

A Política Aleatória, ao fazer escolhas aleatórias, não tem nenhuma informação sobre a distribuição subjacente dos dados. Portanto, é esperado que a Distância de Benford (WC) resultante de suas escolhas esteja mais distante da distribuição teórica de Benford do que a Distância de Benford (WC) resultante das escolhas feitas pelo algoritmo de Aprendizado por Reforço.Se essa expectativa for confirmada, pode-se inferir que as escolhas do algoritmo de Aprendizado por Reforço refletem seu aprendizado sobre a aproximação de uma função que faz essa seleção.

Em outras palavras, o algoritmo de Aprendizado por Reforço está aprendendo a tomar decisões que resultam em uma Distância de Benford (WC) mais próxima da distribuição teórica de Benford, que é o objetivo desejado. Portanto, a Política Aleatória serve como uma métrica de comparação útil para avaliar o progresso e a eficácia do algoritmo de Aprendizado por Reforço.

3.3.2 Agente DQN

Os autores Mnih et al. (2015) apresentam uma nova abordagem para o Aprendizado por Reforço chamada *Deep Q-Network* (DQN). A DQN combina o AR com redes neurais profundas para criar um Agente capaz de aprender a jogar uma variedade de jogos de vídeo a um nível humano.

A técnica de *Deep Q-Network* é baseada em uma rede neural profunda para

aproximar a função de valor-Q, que é uma função que estima o valor esperado de recompensa futura para cada ação possível em um dado estado. A entrada para a rede neural é o estado atual do ambiente (por exemplo, a tela do jogo de vídeo), e a saída é um vetor de valores-Q estimados para cada ação possível.

Os pesos da rede neural são atualizados usando uma variante do algoritmo de Q-learning, que é um algoritmo de Aprendizado por Reforço baseado em diferença temporal. A atualização é feita de forma a minimizar a diferença entre os valores-Q estimados pela rede neural e os valores-Q alvo, que são calculados com base na recompensa recebida e no valor-Q estimado para o próximo estado.

Seja $Q(s, a; \theta)$ a função de valor-Q estimada pela rede neural com pesos θ para a ação a no estado s. Seja r a recompensa recebida após tomar a ação a no estado s, e s' o próximo estado.

Seja $y = r + \gamma \max_{a'} Q(s', a'; \theta)$ o valor-Q alvo, onde γ é o fator de desconto que determina a importância das recompensas futuras. Os pesos da rede neural são atualizados para minimizar a função de perda $(y - Q(s, a; \theta))^2$.

No algoritmo DQN (*Deep Q-Network*) (MNIH et al., 2013), a função Q-valor $Q^*(s, a)$ é aproximada por uma rede neural profunda. A atualização do Q-valor segue a equação do algoritmo *Q-learning*. A função $Q^*(s, a)$ representa o valor esperado do retorno acumulado ao tomar uma ação a em um estado s seguindo uma política ótima.

A equação para atualizar o Q-valor usando a função de recompensa imediata $R(s, a)$ e o Q-valor do próximo estado s' é:

$$Q^*(s, a) = Q(s, a) + \alpha \left[R(s, a) + \gamma \max_{a'} Q(s', a') - Q(s, a) \right]$$

Na qual: $Q(s, a)$ é o Q-valor atual para o estado s e a ação a. α é a taxa de aprendizado que controla a rapidez com que os valores Q são atualizados. $R(s, a)$ é a recompensa imediata obtida ao tomar a ação a no estado s. γ é o fator de desconto que controla o peso dado às recompensas futuras em relação às recompensas imediatas. $\max_{a'} Q(s', a')$ é o valor máximo dos Q-valores para todas as ações possíveis no próximo estado s'.

A ideia do algoritmo DQN é treinar a rede neural para aproximar os Q-valores de forma a minimizar o erro entre os Q-valores estimados e os Q-valores alvo calculados usando a equação acima. Isso é feito usando um processo de aprendizado de reforço que utiliza amostras de experiência armazenadas em um *buffer* de *replay* (MNIH et al., 2013).

Portanto , a expressão completa para $Q^*(s, a)$ na equação é:

$$Q^*(s, a) = Q(s, a) + \alpha \left[R(s, a) + \gamma \max_{a'} Q(s', a') - Q(s, a) \right]$$

Com isso, tem-se os elementos necessário para detalhar o processo de treinamento.

3.3.3 O processo de treinamento

A arquitetura da rede neural utilizada neste trabalho é do tipo sequencial ou *Sequential*, ou seja, uma sequência linear de camadas. Essa arquitetura é utilizada como modelo de referência (*baseline*) no contexto do problema em questão, conforme demonstrado na Tabela 4.

Tabela 4 – Sequential Model Architecture

Layer (type)	**Output Shape**	**Param #**
Dense :: Batch Normalization :: Dropout	1024	79.872
Dense :: Batch Normalization :: Dropout	512	524.800
Dense :: Batch Normalization :: Dropout	256	131.328
Dense :: Batch Normalization :: Dropout	128	32.896
Dense :: Batch Normalization :: Dropout	64	8.256
Dense :: Batch Normalization :: Dropout	32	2.080
Dense :: Batch Normalization :: Dropout	16	528
Dense :: Batch Normalization :: Dropout	4	68
Dense :: Batch Normalization :: Dropout	2	10

Fonte: Elaborado pelo autor

A rede neural utilizada neste trabalho é composta por várias camadas densas (*fully connected*), onde cada neurônio de uma camada está conectado a todos os neurônios da camada seguinte. A primeira camada possui 1024 neurônios e recebe como entrada um vetor de tamanho *state_size*, ou seja, considerando o tamanho das variáveis de estado que são compostas tanto por variáveis do ambiente quanto por variáveis históricas. Essa camada é seguida por uma camada de *Batch Normalization*, que normaliza os valores de ativação dos neurônios para melhorar o treinamento. Em seguida, é aplicada uma camada de *Dropout*, que desativa aleatoriamente uma fração de neurônios durante o treinamento para evitar overfitting.

Essa sequência de camadas densas, *Batch Normalization* e *Dropout* é repetida em várias camadas, com diminuição progressiva no número de neurônios. A última camada possui *action_size* neurônios, que corresponde ao número de ações possíveis do Agente. A função de ativação utilizada nas camadas intermediárias é a função *swish*, que é uma função de ativação não linear. A função de ativação da camada de saída é definida por

um parâmetro chamado *final_layer_activation*, cujo valor é registrado e armazenado no registro de experimentos do MLflow e Tensorboard [4].

O otimizador utilizado é o Adam, que é um algoritmo de otimização amplamente utilizado em redes neurais. A função de perda (loss) utilizada é a *Mean Absolute Error* (MAE), que é uma medida da diferença média entre as previsões do modelo e os valores reais. Métricas adicionais, como o erro quadrático médio (MSE), erro absoluto médio (MAE) e acurácia, também são calculadas durante o treinamento.

Essa arquitetura permite a construção de um modelo de aprendizado profundo que é utilizado no contexto específico do problema em questão, com a capacidade de realizar previsões e tomar decisões com base nas informações fornecidas. A cada passo do experimento, o Agente toma uma decisão de aprovar ou rejeitar o pagamento da indenização com base em diversas variáveis do ambiente, incluindo o percentual de orçamento utilizado, o percentual aprovado e as variáveis WoC (Without Current) e WC (With Current).

As variáveis WoC representam o cenário sem considerar a informação atual, ou seja, apenas com base em informações históricas e sem levar em conta os dados mais recentes. Já as variáveis WC consideram a informação atual, incorporando os dados mais recentes na tomada de decisão.

Ao utilizar tanto as variáveis históricas quanto as variáveis que consideram a informação atual, o Agente busca aproveitar o conhecimento adquirido a partir de padrões passados, ao mesmo tempo em que leva em conta a situação atual para tomar decisões mais precisas e informadas. Dessa forma, o Agente utiliza uma combinação de informações passadas e presentes para calcular a recompensa com base na distância da distribuição teórica de Benford e aprimorar suas decisões ao longo do tempo.

3.4 Experimentos

Neste trabalho, dois principais experimentos são propostos: no primeiro, uma política aleatória para validar a estabilidade e consistência do ambiente; no segundo, o Agente DQN será treinado com uma configuração básica descrita nesta seção, com a qual pode-se comparar os resultados aleatórios. Conforme indicado pela Tabela 5, serão testados valores de Orçamento *(Budget)* que variam de R$ 1 a R$15 Milhões, sempre com 10 episódios para manter a comparação com mesma base, ou seja, os Agentes terão a mesma chance para performar a tarefa, considerando os mesmos valores de orçamento e episódios para os dois casos.

[4] Pacotes para rastreamento e análise de experimentos

Tabela 5 – Configuração Básica dos Experimentos

Variável	Valor
`BUDGET(R$)`	1M, 5M, 10M, 15M
`EPISODES`	10

Fonte: Elaborado pelo autor

3.5 Política Aleatória

Este experimento consiste em utilizar uma política de decisão aleatória no ambiente de Aprendizado por Reforço. Ou seja, a Função de Ação 3.3 será guiada por uma seleção aleatória com 50% [5] de chance a partir do Espaço de Ações. Assim, o Agente toma suas decisões de forma aleatória, sem levar em consideração nenhuma informação específica do ambiente ou do objetivo a ser alcançado.

O objetivo desse experimento é avaliar o comportamento e o desempenho do Agente sob uma política aleatória. Isso permite compreender como o Agente interage com o ambiente e qual é o resultado obtido sem a influência de uma estratégia específica.

3.5.1 O processo de seleção amostral

Como cada episódio é limitado pelo valor de orçamento ou *budget*, sendo que ele deve encerrar-se uma vez que o valor mínimo a ser indenizado na amostra é maior do que o potencial valor ainda no saldo da Carteira C_p. Assim, é esperado que número total de observações selecionados para o treinamento do algoritmo seja inferior ou igual ao total disponível nessa amostra:

$$card(Obs_{train}) \geqslant card(Obs_e) \text{ para todo } e <= e* \in \mathbf{N} \text{ para um } e* \text{ fixo} \tag{3.8}$$

A seleção aleatória com 50% de chances a partir da distribuição binomial pode ser realizada utilizando uma função de geração de números aleatórios que segue essa distribuição. A distribuição binomial modela o número de sucessos em uma série de ensaios independentes, cada um com uma probabilidade constante de sucesso.

[5] A escolha de 50% representa o caso base, mas outros valores podem ser testados em oportunidades futuras.

3.6 Experimento com Agente DQN

O modelo utilizado para o experimento é denominado Seguro Agrícola Brasileiro, computacionalmente codificado como *br_crop_insurance* e nele, o número de episódios definido para o treinamento do Agente é 10 de acordo com as configurações básicas já descritas nesta seção. Conforme descrito na Tabela 3.6, para o armazenamento das experiências do Agente na memória de *replay* é estabelecido um tamanho mínimo de memória de 200, sendo que a atualização da rede neural alvo ocorre a cada 10 episódios.

Tabela 6 – Configurações do Experimento DQN

Variável	**Valor**
BUDGET(Milhões R$)	1, 5, 10, 15
EPISODES	10
MIN_MEMORY	200
UPDATE_TARGET	10
MEMORY_SIZE	500
BATCH_SIZE	50
LEARNING_RATE	0.005
DROP_OUT_RATE	0.9
EPSILON_DECAY	0.7
MIN_EPSILON	0.01

Fonte: Elaborado pelo autor

Nesse ponto, é mencionado o armazenamento das experiências do Agente em uma memória de *replay*. Essa memória é uma estrutura de dados que armazena informações sobre as ações tomadas pelo Agente, as observações do ambiente e as recompensas recebidas em cada etapa. A definição de um tamanho mínimo de memória de 200 significa que o Agente só começará a utilizar a memória de *replay* após ter acumulado pelo menos 200 experiências. Isso é importante para garantir que haja uma quantidade mínima de dados disponíveis para o Agente aprender e tomar decisões baseadas em experiências passadas.

Além disso, a atualização da rede neural alvo ocorre a cada 10 episódios. A rede neural alvo é uma cópia da rede neural principal do Agente, que é usada para estimar os valores de Q durante o treinamento. A cada 10 episódios, a rede neural alvo é atualizada com os parâmetros da rede neural principal, permitindo que o Agente aprenda com uma versão mais estável da rede neural.

Essas estratégias de armazenamento e atualização da memória de replay e da

rede neural alvo são utilizadas para melhorar o desempenho do Agente e auxiliá-lo no processo de aprendizado e tomada de decisões. A capacidade total da memória de replay é definida como 500, e o tamanho dos lotes de treinamento é fixado em 50.

Os hiperparâmetros relacionados ao aprendizado do Agente incluem uma taxa de aprendizado de 0.005 e uma taxa de desativação (dropout rate) de 0.9. Além disso, é aplicada uma taxa de decaimento do epsilon (exploração) igual a 0.7, garantindo uma redução gradual na exploração ao longo do treinamento. Um valor mínimo de epsilon é definido como 0.01.

Por fim, é estabelecido um critério de parada denominado *early stopping*, com base na métrica de perda (*loss*). O treinamento será interrompido se a perda não apresentar melhoria após 5 iterações consecutivas, utilizando o modo *min* (busca de minimização) como critério de comparação. Essa estratégia visa evitar o treinamento excessivo e reduzir o risco de *overfitting* [6].

3.6.1 Avaliação do Agente DQN

A Tabela 7 introduz métricas de avaliação fundamentais comumente utilizadas para avaliar o desempenho de modelos de classificação. Essas métricas desempenham um papel crucial na avaliação da eficácia de modelos preditivos em diferentes domínios.

Tabela 7 – Definições de Acurácia, Precisão, Revocação e F1 Score

Métrica	**Definição**
Acurácia	$\frac{\text{Número de previsões corretas}}{\text{Número total de previsões}}$
Precisão	$\frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Positivos}}$
Revocação	$\frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}}$
F1 Score	$2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$

Fonte: Elaborado pelo autor.

As métricas abrangem a Acurácia, que mensura a proporção de instâncias corretamente previstas; a Precisão, que avalia a Precisão das previsões positivas; o Revocação, que quantifica a capacidade do modelo de identificar corretamente todas as instâncias

[6] Situação indesejável para um algoritmo que pretende generalizar o aprendizado de uma tarefa.

positivas; e o F1 Score, que oferece uma avaliação equilibrada ao combinar Precisão e Revocação. Cada métrica é definida matematicamente, garantindo uma compreensão clara de seu cálculo e interpretação, proporcionando meios esclarecedores de avaliar modelos de classificação.

Essas métricas são úteis para avaliar diferentes aspectos do desempenho do modelo. Por exemplo, a acurácia indica a taxa geral de sucesso das predições, enquanto a revocação pode ser importante quando se deseja identificar corretamente os casos positivos, mesmo que isso implique em alguns falsos positivos. A precisão, por sua vez, indica a proporção de casos positivos corretamente identificados em relação aos casos identificados como positivos.

Dada a importância dos modelos de classificação em uma variedade de aplicações, compreender essas métricas de avaliação é essencial para tomar decisões bem fundamentadas sobre o desempenho do modelo. Essas métricas permitem quantificar a qualidade das previsões, ponderar a Precisão em relação a revocação e a eficácia geral do modelo em contextos distintos.

3.7 Recursos

Com as variáveis e o *framework* de modelagem disponíveis, serão utilizados computadores em nuvem para execução do treinamento, logs e persistência dos dados finais versionados. Assim, a linguagem Python 3.10 (ROSSUM; DRAKE, 2009) e suas bibliotecas Pandas (MCKINNEY et al., 2010), OpenAI (BROCKMAN et al., 2016) e Numpy (HARRIS et al., 2020) serão as principais bibliotecas de software utilizados para a implementação dos algoritmos. Serão versionados o código com a ferramenta Git (CHACON; STRAUB, 2014), o qual será disponibilizado em repositório público a ser definido com a finalidade de facilitar o acesso ao trabalho.

3.8 Dados

A principal fonte de dados deste trabalho é pública e está disponível no site do Programa de Seguro Rural (MAPA, 2021), no qual tem-se dados desde o ano de 2006 até 2021, constando de propostas com dados relevantes para a avaliação da indenização como: cultura, localização geográfica e valores de cobertura, subvenção e indenização. As variáveis brevemente descritas abaixo estão detalhadas no Apêndice A [7].

O grupo **Informações de Proposta e Apólice** engloba todos os detalhes relacionados à proposta de seguro e à apólice emitida. Isso inclui datas de proposta e apólice, identificação da proposta, número da apólice, processo SUSEP e períodos de

[7] Tabela 17

vigência. Essas informações são fundamentais para rastrear a linha do tempo e o status de uma apólice de seguro.

O grupo **Informações de Localização** concentra-se em detalhes geográficos. Ele inclui várias formas de latitude e longitude, município e estado da propriedade, e códigos geográficos. Essas informações são vitais para entender a localização exata da propriedade segurada e podem ser usadas para análises geoespaciais.

Informações de Segurado é um grupo que contém detalhes sobre o indivíduo ou entidade que adquiriu o seguro. Isso inclui o nome do segurado, o número do documento e a razão social. Essas informações são úteis para identificar e entender melhor o perfil do segurado.

O grupo **Informações de Produto e Cultura** inclui detalhes sobre o tipo de produto segurado e a cultura ,isso inclui a classificação do produto, a produtividade segurada e estimada, sendo informações são essenciais para entender o objeto do seguro.

Informações de Cobertura e Prêmio é um grupo que contém detalhes sobre a cobertura do seguro e os prêmios associados. Isso inclui o nível de cobertura, a taxa, o prêmio líquido, o limite de garantia e a subvenção federal. Essas informações são cruciais para entender os detalhes financeiros da apólice de seguro.

O grupo **Informações de Evento e Indenização** inclui detalhes sobre qualquer evento que possa ter desencadeado uma indenização e o valor da indenização. Essas informações são importantes para entender a frequência e o impacto dos eventos segurados.

Finalmente, as variáveis do grupo **Outras Informações** possuem detalhes adicionais a respeito do tamanho da área total segurada e o número de animais, no casos de seguros com esse tipo de cobertura. A fim de produzir o treinamento de algoritmos de reforço, será utilizada uma amostra com todas as propostas contratadas no Estado do Paraná [8], considerando todos os períodos disponíveis e limitando-se à cultura da Soja cuja amostra consta de 108.321 observações, com 36 variáveis [9].

Esses são os dados iniciais para que possam ser geradas variáveis, utilizando a combinação de dados já disponíveis nesse mesmo conjunto. Por exemplo, uma variável de interesse pode ser o número de indenizações já recebidas por um mesmo segurado até o período da avaliação da proposta.

Após a geração de novas variáveis, serão abordados diferentes métodos para a aplicação de Aprendizado por Reforço, como modelos baseados em processos de Markov com ou sem restrição (KRASHENINNIKOVA et al., 2019), utilizando ambientes adversários

[8] Esta escolha está baseada nas caraterísticas do estado e sua importância relativa ao número e valores de contratos realizados

[9] VIDE Apêndice A

(MEAD et al., 2018), propondo novas formulações da função de recompensa, inclusive, a partir de abordagens que utilizam a Distribuição de Benford (LU; BORITZ; COVVEY, 2006)(MEAD et al., 2018).

4 Resultados e Discussão

Neste trabalho, o problema da construção de carteiras de indenização é abordado como uma sequência de decisões. A ideia central pode ser resumida com o seguinte exercício: considere uma sequência de números $S = \{s_1, s_2, ..., s_{n*}\}$, onde $s_{n+1} = s_n + r$ para algum número real r, e $n* = |S|$ é o número total de elementos na sequência. O número de possíveis combinações binárias dos elementos na sequência é 2^{n*}, ou seja, todas as possíveis sequências para aprovação ou rejeição de indenizações, montando a carteira de indenização com todas as observações aprovadas. Conforme definido anteriormente: a $P_{\text{observado}}(D_i = d)$ probabilidade observada do i-ésimo dígito ser d, e a $P_{\text{teórico}}(D_i = d)$ probabilidade teórica do i-ésimo dígito ser d de acordo com a Lei de Benford. A diferença de Benford para uma dada observação n é definida como:

$$B(S) = \frac{1}{3} \sum_{i \in \{1,2,3\}} \frac{\max_d |P_{\text{observado}}(D_i = d|S) - P_{\text{teórico}}(D_i = d)|}{||P_{\text{observado}}(D_i = d|S) - P_{\text{teórico}}(D_i = d)||}$$

Sendo $S \in [S_1, S_2, ...S_n]$, ou seja, uma sequência do conjunto de sequências formado pelas 2^{n^*} possíveis sequências. O problema é encontrar o S^* que minimiza $B(S)$. Isso pode ser escrito como:

$$\min_S B(S)$$

Esse é um problema de minimização que pode ser resolvido analiticamente, porém na dinâmica prática dos mercados de seguros, as indenizações são variáveis aleatórias. Ou seja, seu valor não é conhecido até que seja realizado e com isso, não é possível calcular o valor *a priori*.

O ambiente *InsurEnv* proposto no capítulo anterior foi avaliado com o experimento aleatório para estabelecer sua estabilidade e consistência. Ou seja, ao propor a dinâmica de sequência de decisões, é preciso estabelecer se seleções aleatórias levam a resultados igualmente aleatórios. Além disso, esse experimento provê uma linha de base para comparação com algoritmos de Aprendizado por Reforço.

4.0.1 Detalhamento do Escopo

O detalhamento do escopo, por sua vez, define os elementos e variáveis que são destacados nesta pesquisa, delimitando o tema para ter um melhor direcionamento prático na abordagem empírica, além de exigir menor complexidade na coleta e processamento de

dados, os quais foram obtidos do sistema de Dados Abertos do MAPA (MAPA, 2021), conforme descrito anteriormente no Capítulo 3.

O primeiro ano dos dados, 2006, mostra uma frequência de seguros do PSR de 15.655. Quinze anos depois, em 2021, a frequência alcançou seu ponto mais alto, com 214.202 seguros, o que representa um aumento de cerca de 1,268% em relação a 2006. No entanto, a frequência caiu para 119.066 em 2022, o que ainda é um aumento considerável de 660% em relação a 2006.

A frequência média de seguros do PSR ao longo do período de 17 anos foi de 70.782 (aproximadamente), o que indica que a frequência em 2021 foi cerca de 202% maior do que a média do período, enquanto a frequência em 2022 foi cerca de 68% maior que a média.

Como parte do escopo, o recorte amostral utilizado foi o processo de seleção de uma parte representativa da população-alvo para ser estudada. Assim, ele viabiliza a generalização dos resultados, com uma seleção adequada e representante da população em trabalho. Isso aumenta a validade e a aplicabilidade dos resultados, permitindo extrapolar as conclusões para a população como um todo bem como a validação dos resultados pela comunidade científica interessada.

Tabela 8 – Estados e culturas: os *Top* 10

UF	**Cultura**	**Frequência**	**Max-Normal**	**Log**	**Z Score**	**CDF**
PR	soja	264.915	1,000	12,487	17,153	1,000
PR	Milho [1]	112.093	0,423	11.627	7,160	0,780
RS	soja	86.090	0,325	11,363	5,459	0,687
RS	Uva	62.744	0,237	11,047	3,933	0,615
SP	soja	56.334	0,213	10,939	3,514	0,563
PR	Trigo	51.708	0,195	10,853	3,211	0,516
GO	soja	46.835	0,177	10,754	2,893	0,473
MS	soja	38.071	0,144	10,547	2,320	0,434
MG	Café	31.863	0,120	10,369	1,914	0,403
RS	Arroz	30.428	0,115	10,323	1,820	0,376

Fonte: Elaborado pelo autor

Como é possível observar a partir da Tabela 8, a soja no Paraná é a cultura dominante em termos de frequência entre todas as culturas. O milho na segunda safra no Paraná vem em segundo lugar, mas com uma frequência significativamente menor. O

arroz no Rio Grande do Sul tem a menor frequência entre as três culturas comparadas. O Z Score, que é uma medida de quão longe uma observação está da média, é o maior para a soja no Paraná, indicando que a frequência dessa cultura é significativamente maior do que a média das frequências de todas as culturas. O valor CDF, que é a função de distribuição cumulativa, é o maior para a soja no Paraná, indicando que a probabilidade de uma observação aleatória ser menor ou igual à frequência da soja no Paraná é 1 (ou 100%).

Além disso, a soja no Paraná tem uma frequência de 264,915, que é mais do que o dobro da frequência do milho na segunda safra no mesmo estado, que tem uma frequência de 112,093. Isso pode indicar que os agricultores de soja estão mais propensos a buscar seguro para suas lavouras.

Embora o seguro para a cultura de milho (2^{a} safra) também seja adquirido acima da média em comparação com outras culturas, sua frequência é significativamente menor que a do seguro para a soja. O escore Z também cai de 17.153 para 7.160, indicando que a frequência dos seguros para o milho no Paraná está mais próxima da frequência média entre todas as combinações (de todos os outros estados e culturas). O segundo maior valor da distribuição acumulada CDF de 0,780 sugere que cerca de 78% das combinações de estado e cultura têm uma frequência menor ou igual a esta, sendo os seguros para a soja no Paraná responsáveis pelos 22% restantes.

Já os seguros para cobertura da soja no Rio Grande do Sul têm uma frequência de 86,090, que é significativamente menor do que a frequência de seguros para a soja no Paraná. No entanto, ainda é uma das maiores frequências entre todas as culturas e estados listados na tabela, indicando que a soja no Rio Grande do Sul também é uma cultura importante para a qual os agricultores buscam seguro. A Tabela 9 mostra alguns eventos que atingiram a região ordenados pela **Frequência**.

Ao observar a coluna **Frequência**, nota-se a diferença entre a ocorrência de secas no PR (43,648) e RS (13,065). Essa é uma diferença significativa, sugerindo que os agricultores no PR estão quase 3.3 vezes mais sujeitos a riscos de seca do que em RS.

Da mesma forma, a Geada no PR (10,578) é quase 3.3 vezes mais comum do que no RS (3,154). Esta é outra diferença significativa, sugerindo que os agricultores no PR podem precisar de mais cobertura para a geada do que os agricultores no RS.

A ocorrência de eventos específicos pode guiar as estratégias de mitigação de risco. Por exemplo, se uma seguradora sabe que a frequência de geada no PR é 10,578, enquanto a de seca em SP é 7,313, ela poderia trabalhar com os agricultores nessas regiões para implementar práticas preventivas específicas.Os eventos mais frequentes normalmente conduzem a uma precificação mais alta do seguro, pois representam um risco maior.

Por exemplo, se a seguradora está avaliando o risco de seca no MS (com

Tabela 9 – Dados dos eventos preponderantes por estado

UF	Evento	Frequência	Max-Normal	Log	Z Score	CDF
PR	Seca	43.648	1,000	10,684	10,425	1,000
RS	Seca	13.065	0,299	9,478	2,967	0,653
PR	Geada	10.578	0,242	9,267	2,360	0,550
RS	Granizo	9.824	0,225	9,193	2,176	0,466
SP	Seca	7.313	0,168	8,897	1,564	0,388
MS	Seca	6.123	0,140	8,720	1,274	0,330
SC	Granizo	4.391	0,101	8,387	0,851	0,281
GO	Seca	3.351	0,077	8,117	0,598	0,246
RS	Geada	3.154	0,072	8,056	0,550	0,220
PR	Chuva Ex.	2.609	0,060	7,867	0,417	0,195

Fonte: Elaborado pelo autor

frequência 6,123) e SC (não listada para seca, mas com Granizo listado com frequência 4,391), ela provavelmente cobrará prêmios mais altos para seguro contra seca no MS, dada a maior frequência do evento.

A diferença de frequência entre esses dois estados para seus respectivos eventos (seca no MS e Granizo em SC) é de 1,732. Isso sugere que a seguradora poderia cobrar prêmios cerca de 39% mais altos para seguro contra seca no MS do que para seguro contra Granizo em SC, tudo mais constante. O evento **chuva excessiva** aparece apenas uma vez na tabela, especificamente para o estado do PR (Paraná). Este evento tem uma frequência de 2,609, o que significa que é o décimo evento mais comum nesta seleção de dados.

Dado que a geada é 45% mais comum no PR do que a seca em SP, a seguradora poderia investir mais recursos na mitigação de riscos para geada no PR do que na seca em SP. Em comparação com outros eventos no mesmo estado (PR): A chuva excessiva é o terceiro tipo de evento mais frequente no PR, depois da "seca"e da "Geada". No entanto, sua frequência é significativamente menor do que a destes dois primeiros eventos (43,648 e 10,578, respectivamente). Isso sugere que, embora a chuva excessiva seja uma preocupação no PR, não é tão comum quanto a seca ou a geada.

A chuva excessiva tem uma frequência de 2,609, que é cerca de 6% da frequência do evento mais comum (seca no PR). Isso indica que, apesar de ser um evento que ocorre, ele é significativamente menos frequente do que o evento mais comum. Além disso, ela tem um valor Max-Normal de 0.060, um Log de 7.867, um Z Score de 0.417 e um CDF de

0.195. Isso sugere que, comparado com outros eventos, a chuva excessiva é menos comum e tem um valor de Z Score relativamente baixo, indicando que está mais próximo da média dos eventos.

Considerações sobre o seguro: A frequência relativamente baixa de chuva excessiva poderia sugerir que os prêmios de seguro para esse evento específico podem ser mais baixos do que para eventos mais comuns, como a seca ou a geada. No entanto, seria importante levar em conta o potencial de danos de um evento de chuva excessiva. Se a chuva excessiva, quando ocorre, causar danos significativos, o prêmio de seguro pode ser ajustado para cima para refletir esse risco.

Dessa forma, são treinados algoritmos com diferentes configurações de *Budget* ou Orçamento, e testados três valores distintos: R$500 mil, R$ 1 milhão e R$1,5 milhão.

4.1 Seleção da Política Aleatória

A seguir, a Tabela 10 traz a distribuição da seleção para os diferentes limites de orçamento da configuração deste experimento, permitindo verificar a ***seleção aleatória*** dos casos conforme demonstrado nas colunas A (Aprovado) representando os aprovados e R (Reprovado) os reprovados.

Tabela 10 – Amostragem por Orçamento - Política Aleatória

Orçamento(R$)	**R**	**A**	**R + A**	**(%)R**	**(%)A**
1M	191	157	348	54,89	45,11
5M	850	825	1675	50,75	49,25
10M	1698	1654	3352	50,66	49,34
15M	2636	2521	5157	51,12	48,88

Fonte: Elaborado pelo autor.

A partir dessa seleção como um modelo de base, foram calculadas as métricas de performance de acordo com o esperado para o agente aleatório e são utilizadas para comparação com o próximo experimento.

Na execução das escolhas aleatórias, o agente não utiliza as variáveis de estado para tomada de decisão. Portanto, o escopo da análise de importância de variáveis é limitado a verificação da aleatoriedade na seleção das amostras dos episódios resultantes em comparação com a amostra original de treinamento.

4.2 Seleção do Agente DQN

Em relação à seleção aleatória, o Agente DQN apresentou uma taxa de reprovação relativamente maior, chegando a quase 1.5 vezes no caso do valor de orçamento R$5 Mi, a qual possui a maior taxa de reprovação entre as taxas apresentadas na Tabela 11. Com isso, o treinamento desse algoritmo produziu a seleção de casos para as diferentes carteiras de indenização, para cada um dos orçamentos, na Tabela 11:

Tabela 11 – Amostragem por Orçamento - Agente DQN

Orçamento(R$)	**R**	**A**	**R + A**	**(%)R**	**(%)A**
5M	44	29	59	22,00	78,00
1M	2067	663	2730	76,00	24,00
5M	8928	2440	11368	79,00	21,00
10M	3046	2194	5240	58,00	42,00

Fonte: Elaborado pelo autor

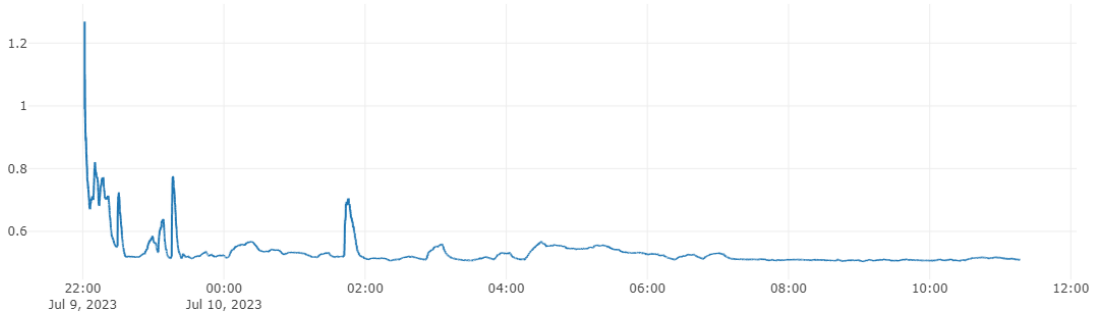
Nesse momento, obteve-se a resposta buscada de interesse neste trabalho: é possível treinar modelos de Aprendizado por Reforço para a tarefa de construção de carteiras de indenização, diminuindo o risco de fraude medido pela distância de Benford apresentados em detalhe no Capítulo 3.

Para inferir o sucesso do aprendizado pelo algoritmo, basta que o algoritmo proposto exiba tanto uma perda decrescente ao longo do tempo quanto um aumento na recompensa acumulada. Esses critérios devem permitir que o Agente DQN supere o modelo de base aleatória em diferentes parâmetros de orçamento, conforme descrito nas Tabelas 10 e 11.

A rede neural subjacente ao Agente DQN demonstrou convergência para um patamar de perda mais baixo, evidenciada pela diminuição da perda ao longo dos episódios. Inicialmente, a perda se encontra próxima de 1.2. No entanto, essa métrica diminui para valores abaixo de 0.5 a partir do 2º minuto de treinamento em diante, conforme pode ser observado na Figura 6.

Para compreender os experimentos de forma detalhada, será verificado como cada agente "experimentou"esse ambiente. Inicialmente, o número de passos pode trazer uma indicativa do quanto os agentes foram expostos ao ambiente, ou seja, quantas experiências e diferentes solicitações de indenizações já foram avaliadas por ele.

Figura 6 – Função de Perda (*Loss*) por Tempo (s)



Fonte: FONTE: Elaborado pelo autor utilizando MLFlow

4.3 Comparando as seleções

As Tabelas 10 e 11 apresentadas mostram a amostragem por orçamento para duas políticas distintas: Política Aleatória e Agente DQN. Cada tabela contém informações sobre o número de casos que foram selecionados pela política (R), o número de casos que foram rejeitados (A), o total de casos considerados na amostra (R + A), a proporção de casos selecionados em relação ao total (em %R) e a proporção de casos rejeitados em relação ao total (em %A).

Comparativamente, podem-se notar algumas diferenças significativas nas proporções de casos selecionados e rejeitados em relação aos orçamentos. Na **Política Aleatória**, os casos selecionados e rejeitados estão distribuídos de forma mais equilibrada, com proporções aproximadamente iguais, em torno de 50%. Por outro lado, nos dados relativos ao experimento com o **Agente DQN**, as proporções são mais desiguais, com uma maior predominância de casos selecionados (mais de 75%) em relação aos rejeitados (menos de 25%), especialmente para orçamentos menores.

Esses resultados sugerem que o **Agente DQN** possui uma tendência mais forte para selecionar casos em relação à **Política Aleatória**, o que pode indicar uma maior eficácia do agente em identificar padrões relevantes para a construção de carteiras de indenização. Entretanto, é importante lembrar que essas conclusões são baseadas na análise da amostragem e, para uma avaliação mais completa e robusta, será necessário realizar uma análise mais aprofundada dos resultados obtidos pelos agentes em relação às métricas de desempenho, tais como acurácia, precisão, revocação (*recall*) e F1-score, entre outras.

4.3.1 Avaliação do Número de Passos

Os resultados são divididos em quatro seções: Máximo (*Passos*), Média (*Passos*), Desvio Padrão *(Passos)* e Contagem (*Passos*). Na seção Máximo (*Passos*), são exibidos os valores máximos alcançados em cada experimento para o número de Passos. Por exemplo, no experimento com orçamento de 1M, o método DQN alcançou um máximo de 176 Passos, enquanto o método Aleatório alcançou um máximo de 53 Passos. Na seção Média (*Passos*), são apresentadas as médias dos números de Passos em cada experimento. Por exemplo, no experimento com orçamento de 5M, o método DQN teve uma média de 651 Passos, enquanto o método Aleatório teve uma média de 87 Passos.

Nota-se que o fato dos passos realizados pelo método DQN terem sido superiores aos passos realizados pela Política Aleatória, evidenciando que o Agente DQN consegue avaliar e explorar o ambiente adquirindo muito mais experiência do que as escolhas aleatórias permitiriam. Além disso, a média de passos realizados pelo Agente DQN é quase o dobro do que a obtida com a **Política Aleatória**.

A seção Desvio Padrão (Passos) mostra os desvios padrão dos números de Passos em cada experimento, indicando a variabilidade dos resultados. Por exemplo, no experimento com orçamento de 10M, o método DQN teve um desvio padrão de 161, enquanto o método Aleatório teve um desvio padrão de 101. Por fim, na seção Contagem (Passos), são apresentados o total de Passos contabilizados em cada experimento. Por exemplo, no experimento com orçamento de 1M, o método DQN teve um total de 2.730 Passos, enquanto o método Aleatório teve um total de 348 Passos.

A Tabela 12 apresenta os resultados experimentais comparando o desempenho do Agente DQN (Deep Q-Network) com um agente aleatório. Os dados incluem o número máximo de passos, média de passos, desvio padrão de passos e o tamanho da amostra n de cada experimento.

Essa tabela permite uma análise comparativa dos resultados entre os métodos DQN e Aleatório em diferentes cenários de orçamento, fornecendo informações sobre o desempenho máximo, médio, variabilidade e contagem de Passos em cada experimento.

O número de passos de um determinado episódio revela o quão longe cada agente conseguiu chegar em relação ao número total de amostras disponíveis. Além disso, quanto maior o número de passos mais provável será de acumular mais recompensas, o que significa mais chances de ter um melhor desempenho. Assim, nota-se que o Agente DQN apresenta números máximos e também a média superiores em número de passos em relação à Política Aleatória. No experimento com 1M , o DQN supera o método aleatório em 123 passos, com uma relação DQN/Aleatório de 3,32. Isso indica que o DQN é mais de três vezes melhor do que o método aleatório neste cenário. Para experimentos com 5 milhões, 10 milhões e 15 milhões de passos, o DQN supera o método aleatório em 2007,

Tabela 12 – Resultados Experimentais: DQN vs. Aleatório

Experimento	DQN	Aleatório	Diferença	DQN/Al.
Máximo (Passos)				
1M	176	53	123	3,32
5M	2223	216	2007	10,29
10M	675	405	270	1,67
15M	988	605	383	1,63
Média (Passos)				
1M	61	20	41	3,05
5M	651	87	564	7,48
10M	268	170	98	1,58
15M	402	259	143	1,55
Desvio Padrão (Passos)				
1M	45	13	32	3,46
5M	610	53	557	11,51
10M	161	101	60	1,59
15M	240	151	89	1,59
Contagem (Passos)				
1M	2730	348	2382	7,85
5M	11368	1675	9693	6,79
10M	5240	3352	1888	1,56
15M	5540	5157	383	1,07

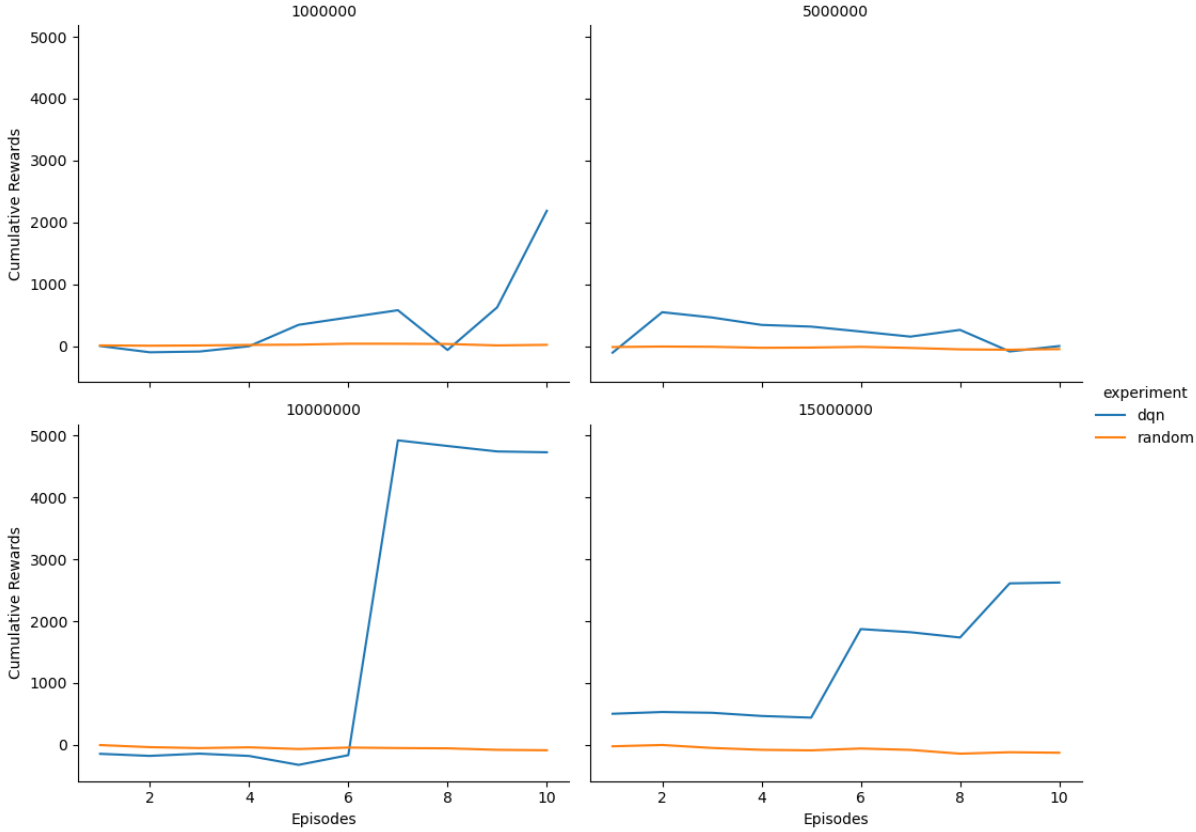
270 e 383 passos, respectivamente, com relações DQN/Aleatório de 10,29, 1,67 e 1,63. O DQN é significativamente superior em todos esses cenários. Além disso, o agente DQN exibiu um desvio padrão maior em relação ao método aleatório em todos os experimentos, o que sugere uma variabilidade maior em seu desempenho. Isso aponta também que o Agente DQN pode ser mais sensível a diferentes condições ou configurações de ambiente, resultando em uma gama mais ampla de resultados possíveis. O DQN exibe um desvio padrão maior em relação ao método aleatório em todos os experimentos. As diferenças variam de 32 a 557 passos, com relações DQN/Aleatório correspondentes de 3,46, 11,51, 1,59 e 1,59. Isso sugere que o DQN pode ser mais volátil em termos de variabilidade nos resultados, especialmente nos experimentos com 5 milhões de passos.

4.3.2 Recompensas

As recompensas acumuladas representam uma métrica essencial para determinar a capacidade de treinamento de algoritmos de aprendizado por reforço nesse ambiente específico. Ao analisar os gráficos apresentados a seguir e observar essa métrica, podemos notar que o Agente DQN treinado obteve um desempenho superior em relação ao experimento de Política Aleatória, conforme esperado.

Até o momento, foram analisados os resultados básicos referentes ao número de passos por episódio em cada experimento. Agora, será realizada uma avaliação das recompensas obtidas durante o experimento, conforme mostrado na Figura 7.

Figura 7 – Recompensas por Episódio e Orçamento



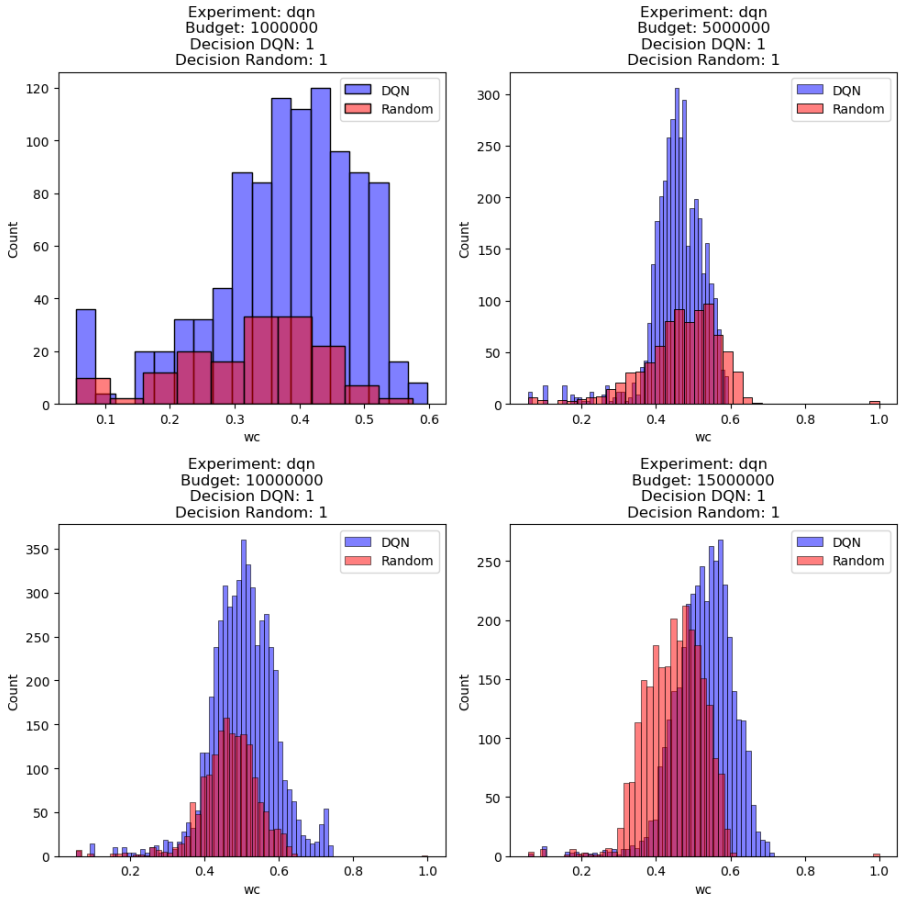
Fonte: Elaborado pelo autor.

Esses resultados evidenciam que é possível treinar algoritmos de aprendizado por reforço com sucesso nesse ambiente, fortalecendo a validade e eficácia do método utilizado. Como é possível observar, o Agente DQN apresenta um desempenho melhor do que a escolha aleatória em todos os possíveis orçamentos iniciais escolhidos. Também é possível verificar que as recompensas aumentaram conforme os orçamentos aumentaram, de forma significativa no caso do orçamento de R$ 10 Milhões em diante.

4.3.3 As variáveis *WC* e *WoC*

Essas variáveis estão diretamente ligadas à ação tomada pelos agentes, uma vez que guardam o valor com ou sem o atual caso sob consideração para aprovação ou rejeição do agente. Como é possível notar na Figura 8, a distribuição da variável WC começa a se diferenciar para os agentes a partir de R$ 10 Milhões, momento no qual a parte da distribuição que representa o Agente DQN passam a assumir valores maiores.

Figura 8 – Distribuição da Variável *WC*



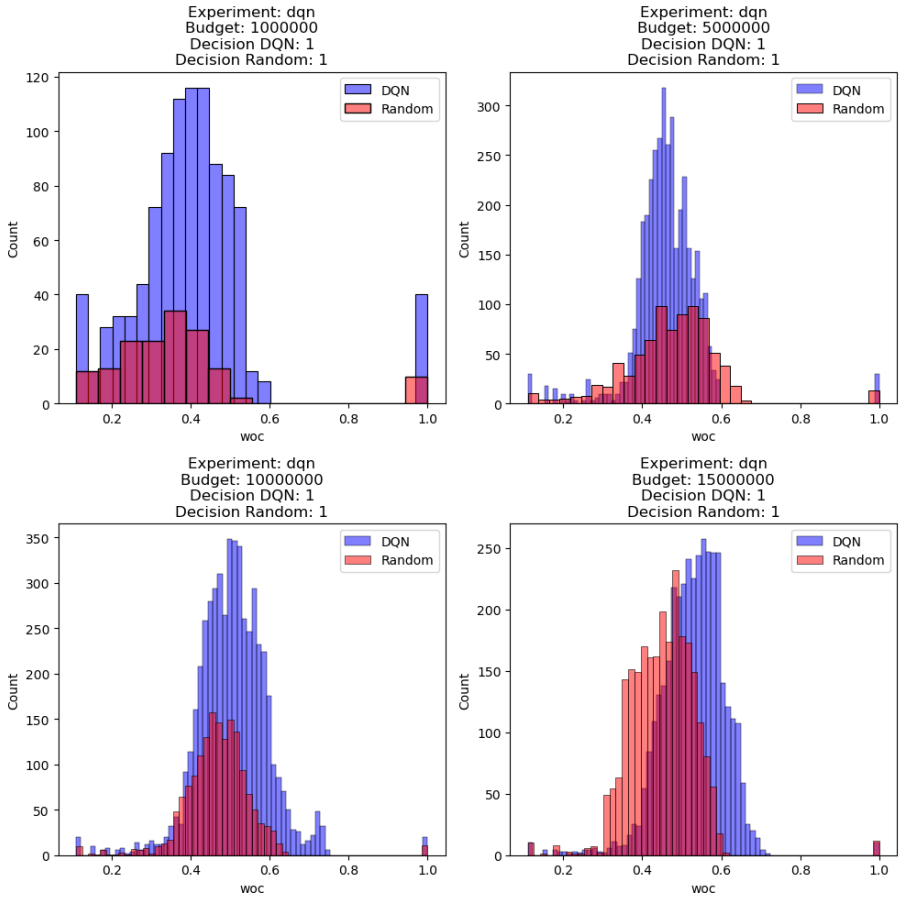
Em resumo, os resultados experimentais mostram que o algoritmo DQN apresentou uma melhoria progressiva no desempenho à medida que o orçamento inicial aumentou, enquanto o algoritmo aleatório não mostrou uma relação clara entre o desempenho e o orçamento inicial.

No caso da Figura 9, de maneira análoga, as diferenças começam a ficar mais evidentes a partir do Orçamento de R$ 10 Milhões, ficando mais clara com o R$ 15 Milhões.

4.3.4 Valor de Indenização

O valor de indenização apresentou um comportamento distinto entre os dois experimentos o que é apresentado na Tabela 13 permite uma análise comparativa em

Figura 9 – Distribuição da Variável *WoC*



Fonte: Elaborado pelo autor.

relação aos valores de indenização máximo, médio, desvio padrão e contagem, fornecendo informações sobre o desempenho e a variabilidade dos métodos em diferentes cenários de orçamento.

A análise detalhada das métricas a seguir desempenha um importante papel na avaliação de algoritmos de aprendizado por reforço, como o DQN, em aplicações do mundo real. Essas métricas são fundamentais para medir o desempenho e a eficácia de tais algoritmos em ambientes complexos e dinâmicos, como o seguro de culturas agrícolas. O valor de indenização máximo, por exemplo, reflete o ponto mais alto que um algoritmo pode atingir em termos de compensação financeira para os agricultores. Compreender como o DQN se compara a uma abordagem aleatória nesse aspecto pode fornecer *insights* valiosos sobre a capacidade do algoritmo de otimizar os recursos disponíveis.

Além disso, a análise das médias dos valores de indenização oferece uma visão mais holística do desempenho geral dos algoritmos. Ela nos permite avaliar se o DQN é capaz de obter compensações financeiras consistentemente ao longo de diferentes cenários de orçamento. O desvio padrão, por sua vez, avalia a variabilidade dos resultados, revelando quão previsível e estável é o desempenho do algoritmo em face de diferentes condições.

Neste contexto, a contagem dos valores de indenização contabilizados adiciona uma dimensão importante à análise, verificando se a quantidade de passos está alinhada com os valores medidos. Por último, esta análise proporciona uma compreensão aprofundada de como o DQN opera os valores financeiros de principal destaque no ambiente. Esta exploração é essencial para o aprimoramento contínuo dos algoritmos de aprendizado por reforço em aplicações do mundo real, e os resultados podem ter implicações significativas para o setor agrícola e além.

Na Tabela 13, são apresentados os resultados experimentais comparando os métodos DQN e Aleatório em diferentes cenários de orçamento. Os valores exibidos são referentes ao valor de indenização máximo, médio, desvio padrão e contagem em cada experimento.

Na seção **Máximo (Valor de Indenização)**, são mostrados os valores máximos alcançados para o valor de indenização em cada experimento. Por exemplo, no experimento com orçamento de 1M, o método DQN alcançou um valor máximo de R$602.123, enquanto o método Aleatório alcançou um valor máximo de R$1.116.572, sendo também o maior valor registrado entre todos os orçamentos, o valor de R$ 1.162.334. Em comparação, a média dos valores máximos do Agente DQN foi de R$771.822, o que demonstra que o DQN, em média, obteve valores de indenização mais baixos do que o método Aleatório. Essa diferença pode ser atribuída à maneira como o Agente DQN explora o ambiente e utiliza seus parâmetros, como o epsilon, que controla a taxa de exploração versus explotação. Uma possível interpretação está na maior ênfase do DQN na exploração, o que pode levá-lo a experimentar mais estados e, portanto, não se concentrar apenas nas ações mais lucrativas. Isso pode resultar em valores máximos mais baixos, mas também pode contribuir para um aprendizado mais abrangente e adaptável em cenários variáveis.

Na seção **Média (Valor de Indenização)**, são apresentadas as médias aritméticas dos valores de indenização em cada experimento. Novamente, observa-se que, em geral, o Agente DQN registra valores médios de indenização inferiores em comparação com a Política Aleatória. A maior diferença é observada no experimento com um orçamento de 1M, onde o Agente DQN apresenta uma média de R$42.952, enquanto a Política Aleatória registra uma média de R$67.043. Esse padrão continua nos outros orçamentos, com o Agente DQN geralmente obtendo médias menores. A diferença diminui um pouco no experimento de 10M, onde o Agente DQN registra uma média de R$44.394 em comparação com a média de R$59.209 da Política Aleatória. Esses resultados sugerem que o Agente DQN tende a gerar valores médios de indenização mais baixos, o que pode ser atribuído ao seu foco na exploração do ambiente e na busca de um equilíbrio entre explorar novas ações e escolher ações com base em experiências passadas. Isso pode ser vantajoso em cenários de mudança, mas também pode resultar em valores médios de indenização mais baixos em comparação com uma abordagem puramente aleatória que não considera o histórico das

Tabela 13 – Resultados Experimentais: DQN vs. Aleatório

Experimento	DQN	Aleatório	Diferença	DQN/Aleatório
Máximo (Valor de Indenização)				
1M	602.123	1.116.572	-514.449	0,5394
5M	902.316	800.000	102.316	1,1279
10M	902.316	1.162.334	-260.018	0,7766
15M	679.536	1.136.938	-457.402	0,5976
Média (Valor de Indenização)				
1M	42.952	67.043	-24.091	0,6403
5M	43.057	62.172	-19.115	0,6925
10M	44.394	59.209	-14.815	0,7505
15M	41.957	59.827	-17.870	0,7020
Desvio Padrão (Valor de Indenização)				
1M	59.666	106.674	-47.008	0,5601
5M	61.257	80.173	-18.916	0,7623
10M	63.666	81.624	-17.958	0,7804
15M	57.231	82.076	-24.845	0,6967
Contagem (Valor de Indenização)				
1M	2.730	348	2.382	7,8276
5M	11.368	1.675	9.693	6,7764
10M	5.240	3.352	1.888	1,5642
15M	5.540	5.157	383	1,0730

ações.

Na seção **Desvio Padrão (Valor de Indenização)**, tem-se os desvios padrão dos valores de indenização em cada experimento, indicando a variabilidade dos resultados. Por exemplo, no experimento com um orçamento de 10M, o método DQN apresentou um desvio padrão de 63.666, enquanto o método Aleatório registrou um desvio padrão de 81.624. Contudo, a maior diferença observada foi de R$-47.008 em relação ao método Aleatório, demonstrando que o DQN apresenta menor variabilidade, representando quase metade do valor registrado para o método Aleatório.

Em resumo, o DQN demonstrou maior consistência e estabilidade nos valores

de indenização em comparação com o método Aleatório, o que é um indicativo positivo de sua eficácia.

No experimento de 1M, o algoritmo DQN produziu uma diferença significativa em relação ao método Aleatório, com uma diferença positiva de R$514.449. Isso indica que o DQN superou o método Aleatório em termos do valor máximo de indenização. Em 5M, o DQN também superou o método Aleatório, mas com uma diferença ainda mais acentuada de R$1.127.9. Neste cenário, o DQN teve um desempenho muito superior em relação ao Aleatório.

No experimento de 10M, embora o DQN ainda tenha superado o método Aleatório, a diferença positiva diminuiu para R$260.018. Isso sugere que, à medida que o valor de indenização aumentou, o DQN teve um desempenho relativamente menor em comparação com o Aleatório. O cenário de 15M também demonstrou uma diferença positiva, mas menor, de R$457.402. Neste caso, o DQN ainda superou o Aleatório, mas com uma diferença menos acentuada em comparação com o experimento de 5M.

Por fim, na seção **Contagem (Valor de Indenização)**, são apresentados o total de valores de indenização contabilizados em cada experimento, conforme apresentado também na Tabela 12. Com isso, fica verificada a consistência entre a quantidade de passos e os valores medidos nesses passos ao longo dos dez episódios testados.

4.3.5 Evento Preponderante

Esse evento é essencial para a aplicação prática dos modelos de aprendizado por reforço na área de seguros, pois está diretamente relacionado à identificação dos casos mais relevantes e representativos para a análise de risco. Esta comparação entre os experimentos é importante porque ele representa uma métrica que indica o desempenho do agente em selecionar os casos corretos para a construção de carteiras de indenização.

A Tabela 14 permite avaliar a distribuição de eventos preponderantes ao longo das diferentes amostras para os orçamentos selecionados. Como é possível observar, os eventos estão distribuídos de forma similar entre as diferentes amostragens.

Ao comparar o **Evento Preponderante** entre os experimentos, podemos verificar se o Agente DQN consegue selecionar um maior número de casos preponderantes em relação à política aleatória. Isso seria um indicativo de que o agente está aprendendo a identificar padrões importantes nos dados e a tomar decisões mais eficazes na construção das carteiras de indenização. Além disso, essa comparação nos permite avaliar o desempenho geral dos modelos de AR em relação a uma abordagem aleatória, o que é fundamental para entender se esses modelos têm potencial para serem aplicados na prática para melhorar a seleção de carteiras de indenização e mitigar o risco de fraudes.

Portanto, comparar o **Evento Preponderante** entre os experimentos nos

Tabela 14 – Número de propostas por evento preponderante

Evento	**(%)5M**	**(%)10M**	**(%)15M**
SECA	80,54	79,09	80,22
CHUVA EXCESSIVA	10,09	9,28	10,08
GRANIZO	5,13	7,22	5,62
INUNDAÇÃO/TROMBA D´ÁGUA	3,10	3,37	3,01
VARIAÇÃO EXCESSIVA DE TEMPERATURA	0,54	0,24	0,39
DEMAIS CAUSAS	0,30	0,48	0,25
GEADA	0,18	0,06	0,16
VENTOS FORTES/FRIOS	NaN	0,12	0,16
INCÊNDIO	0,06	0,12	0,10
RAIO	0,06	NaN	0,02
REPLANTIO		0,03	

Fonte: Elaborado pelo autor,

fornece elementos analíticos sobre a capacidade do Agente DQN em selecionar casos relevantes e contribui para a validação e interpretação dos resultados obtidos pelos modelos de AR no contexto específico da formação de carteiras de indenização em seguros.

A Tabela 15 fornece informações sobre a porcentagem e contagem de eventos nos conjuntos aleatório e de treinamento, bem como a diferença percentual entre eles. O evento com a maior diferença percentual entre os conjuntos aleatório e de treinamento é seca, com uma diferença de 12.5% na representação. Outros eventos que apresentam diferenças significativas entre os conjuntos são **chuva excessiva**, com uma diferença de 12.3%, e granizo, com uma diferença de 0.4

As variáveis de estado serão examinadas para diferenciar o comportamento do agente aleatório, cobrindo as variáveis chamadas de *Variáveis do Episódio* neste trabalho e que foram apresentadas na Metodologia. As variáveis históricas, contudo, não são apresentadas no detalhe por não trazerem informações relevantes para a comparação dos agentes nesse momento. [2] As principais dimensões do problema que cabem ser destacadas são: o valor de indenização, pela sua importância na função de recompensa, e os resultados das recompensas acumuladas por decisão.

[2] Contudo, o Apêndice A com o sumário explicativo para todas as variáveis.

Tabela 15 – Amostragem por Eventos Preponderantes

Evento	**(%)Rnd**	**(n)Rnd**	**(%)Train**	**(n)Train**	**(%) Diff**
CHUVA EXCESSIVA	9,8	259,5	22,1	1742	-12,3
INUNDAÇÃO TROMBA D´ÁGUA	3,2	83,2	3,7	290	-0,5
DEMAIS CAUSAS	0,3	8,8	0,5	38	-0,2
VARIAÇÃO DE TEMPERATURA	0,5	12,3	0,7	56	-0,2
RAIO	0,0	1,0	0,1	7	-0,1
INCÊNDIO	0,1	3,3	0,1	5	0,0
REPLANTIO	0,0	1,0	0,0	1	0,0
GEADA	0,2	4,3	0,1	4	0,1
VENTOS (FORTES e FRIOS)	0,2	4,3	0,1	6	0,1
GRANIZO	6,1	159,8	5,7	450	0,4
SECA	79,6	2102,8	67,1	5297	12,5

Fonte: Elaborado pelo autor.

4.3.6 Avaliação do Modelo

Ainda que as métricas tradicionais de avaliação de modelos não sejam suficientes para dar conta da complexidade de implementação dos algoritmos de AR, elas trazem importantes informações e que são validas para complementar o estudo da performance do Agente DQN nesta pesquisa.

Tabela 16 – Resultados Experimentais: DQN

Orçamento(R$)	**Acurácia**	**Precisão**	**Recall**	**F1**
1 M	0,37	0,38	0,05	0,08
5 M	0,45	0,42	0,20	0,27
10 M	0,35	0,41	0,17	0,24
15 M	0,41	0,43	0,15	0,23

Fonte: Elaborado pelo autor

A acurácia mede a proporção de predições corretas feitas pelo modelo em relação ao total de exemplos. Por exemplo, quando o orçamento é de 1 milhão de reais, o modelo

DQN alcançou uma acurácia de 0.37, o que significa que ele acertou aproximadamente 37% das predições. Ou seja, o algoritmo conseguiu acertar os passos para diminuir a distância da distribuição teórica de Benford nessa proporção.

A precisão é a proporção de verdadeiros positivos (casos corretamente identificados como positivos) em relação ao total de casos identificados como positivos pelo modelo. No caso do orçamento de 1 milhão, a precisão foi de 0.38, indicando que cerca de 38% dos casos identificados como positivos pelo modelo eram realmente positivos.

A revocação mostra uma melhoria considerável quando o orçamento aumenta de 1 milhão (0.05) para 5 milhões (0.20). No entanto, a revocação diminui novamente nos orçamentos de 10 milhões (0.17) e 15 milhões (0.15). Essa variação sugere que o aumento do orçamento pode ter um impacto positivo inicial na capacidade do modelo de identificar casos positivos, mas há outros fatores em jogo que podem diminuir essa capacidade à medida que o orçamento continua a aumentar.

O F1-score combina a precisão e a revocação em uma única medida, trazendo uma visão geral do desempenho do modelo, sendo calculado como a média harmônica entre a precisão e a revocação. Para o orçamento de 1 milhão, o F1-score foi de 0.08. Comparando os diferentes orçamentos, o orçamento de 5 milhões apresenta o maior F1-score (0.27), indicando um equilíbrio entre precisão e recall. No entanto, os outros orçamentos têm F1-scores mais baixos, sugerindo um desempenho geral inferior em relação ao orçamento de 5 milhões.

Além disso, é importante observar que o desempenho do modelo pode variar de acordo com o orçamento. Por exemplo, a revocação apresentou uma melhoria significativa de 0.05 para 0.20 quando o orçamento aumentou de 1 milhão para 5 milhões. Isso pode indicar que o modelo se beneficia de um orçamento maior para realizar predições mais precisas. No entanto, outros fatores também devem ser considerados na interpretação dos resultados, como a disponibilidade de dados e o contexto específico da tarefa.

É importante ressaltar que, embora os experimentos tenham sido conduzidos com o total de dez episódios, modelos de AR podem apresentar resultados muito mais atrativos com centenas ou milhares de episódios, a depender da estratégia de treinamento que é utilizada. Quando mais episódios, mais oportunidades são dadas ao agente para refinar suas estratégias e explorar o ambiente.

5 Considerações Finais

O objetivo principal do trabalho consistiu em propor e avaliar um ambiente de Aprendizado por Reforço destinado à composição de carteiras de indenização. Para isso, o trabalho utilizou tanto dados históricos quanto a interação do agente com o ambiente. Como resultado desejado do sucesso desse objetivo, o trabalho visou minimizar a possibilidade de fraude, mitigando, portanto, os riscos associados à interação do agente com o ambiente.

Dessa forma, o problema de composição de carteiras com menores riscos de fraude foi modelado como uma sequência de decisões. Nesse modelo, o risco de fraude foi definido como variável teórica e a distância de Benford a variável operacional, a qual permitiu identificar se uma composição está mais ou menos propensa a apresentar divergências em suas indenizações e até potencialmente fraudes.

Assim, o ambiente *InsurEnv* foi elaborado a partir das inspirações obtidas na revisão da literatura bem como elementos de Aprendizado por Reforço. Além disso, o ambiente foi devidamente validado, o que pode ser verificado no Experimento Aleatório apresentado no capítulo anterior. Nele, notou-se que ações aleatórias levam a seleções de carteiras igualmente aleatórias, bem como o resultado das recompensas acumuladas que ficaram estáveis, indicando sucesso na resposta aleatória da função de recompensa.

Para a avaliação dos experimentos, foram estabelecidas as métricas de avaliação dos agentes aleatório e DQN, decompondo as ações para as diferentes dimensões do problema como o Valor da Indenização, *WC* e *Woc*, além da verificação dos principais resultados para a rede neural com métricas clássicas como acurácia, precisão e revocação.

O trabalho explorou a interseção entre diferentes disciplinas, como Economia, Atuária, Psicologia, Matemática, Estatística e Ciência da Computação, a fim de avançar nas aplicações de Aprendizado por Reforço, especialmente em contextos de fraudes. Ao integrar esses conhecimentos, esta pesquisa ofereceu uma abordagem inovadora para a construção de portfólios ou carteiras de indenização de forma a mitigar o risco de fraude, mesmo na ausência da sua definição. Além disso, buscou-se estabelecer as motivações por trás das fraudes e como essas motivações se relacionam com os custos associados a elas.

Ao realizar um exercício teórico, o trabalho destacou a importância de considerar os incentivos e as consequências econômicas das fraudes, incluindo as políticas públicas voltadas para seguros agrícolas no Brasil. Por meio de modelos e algoritmos de Aprendizado por Reforço, utilizando o estado do Paraná e a cultura da soja como escopo amostral, foi possível analisar os diferentes cenários e tomar decisões que visam minimizar os custos decorrentes das fraudes.

Essa abordagem multidisciplinar permitiu uma compreensão mais profunda dos aspectos comportamentais, financeiros e estatísticos relacionados às fraudes. Além disso, os resultados obtidos no trabalho forneceram insights valiosos sobre como lidar com fraudes de forma mais eficaz, tanto em termos de prevenção quanto de detecção.

No entanto, é importante ressaltar que este trabalho possui limitações, como a simplificação da inferência de fraude pela distância de Benford e a dependência de suposições teóricas acerca do impacto de fraudes nos preços de seguros. Portanto, existem oportunidades para pesquisas que explorem, por exemplo, a utilização de múltiplas variáveis de Benford para o problema, expandindo o modelo para uma classificação multi-classes.

Além disso, é possível testar outras abordagens de Aprendizado por Reforço e modelos que possam superar os resultados aqui descritos. Uma oportunidade futura é avaliar as métricas de performance a partir da variável resposta real, trazendo mais um elemento para a construção de recompensas em um esquema de auto-aprendizado e de supervisão.

Para avançar com a metodologia proposta neste trabalho, é possível obter uma complementação da base de dados com um conjunto de variáveis relativas a clima, tipo de solo, relevo, bem com modelos de previsão meteorológica como dados de entrada para o ambiente. Dessa forma, os agentes podem expandir o conhecimento que adquirem no processo de treinamento com mais dimensões de entrada para a rede neural, por exemplo[1].

Cabe ressaltar que diversas outras configurações podem ser testadas, contudo, o objetivo deste trabalho pode ser considerado como cumprido uma vez que, com o escopo delimitado, foi possível realizar o treinamento de agentes de Aprendizado por Reforço mesmo com a assimetria de informação relativa à definição de fraude. Oportunamente, variações nas técnicas e abordagens escolhidas para o agente de aprendizado podem enriquecer e facilitar adoção do ambiente *InsurEnv*.

[1] Dados coletados por sensoriamento remoto, por exemplo, podem trazer imagens da área segurada no período para analise do contrato e potencial indenização.

Referências

ADAMI, A. C. d. O.; OZAKI, V. A. Modelagem estatística dos prêmios do seguro rural. *Revista de Política Agrícola*, v. 21, n. 1, p. 60–75, 2012. Citado na página 21.

AMARAL, F. J. G. d.; BACHA, C. J. C. Subvenções federais dadas à agropecuária brasileira no período de 2003 a 2019. *Revista de Economia e Sociologia Rural*, SciELO Brasil, v. 61, 2022. Citado na página 15.

BABCOCK, B. A. Using cumulative prospect theory to explain anomalous crop insurance coverage choice. *American Journal of Agricultural Economics*, Wiley Online Library, v. 97, n. 5, p. 1371–1384, 2015. Citado na página 25.

BERGER, A.; HILL, T. P. A basic theory of benford's law . *Probability Surveys*, v. 8, p. 1–126, 2011. Citado 3 vezes nas páginas 37, 45 e 46.

BRASIL. Decreto-Lei, *DECRETO-LEI 73, DE 21 DE NOVEMBRO DE 1966. Dispõe sôbre o Sistema Nacional de Privados, regula as operações de seguros e resseguros e dá outras providências*. Brasília, DF, 1966. Incorporado ao Código Civil. Disponível em: <https://www.planalto.gov.br/ccivil\%5F03/decreto-lei/del0073.htm>. Citado na página 21.

_____. Lei, *Lei 10.823, de 19 de dezembro de 2003. Dispõe sobre a subvenção econômica ao prêmio do seguro rural*. Brasília, DF, 2003. Publicada no Diário Oficial [da] República Federativa do Brasil. Disponível em: <http://www.planalto.gov.br/ccivil\%5F03/Leis/2003/L10.823.htm>. Citado 5 vezes nas páginas 12, 15, 18, 21 e 23.

_____. Resolução nº 77, de 22 de setembro de 2020. *Diário Oficial da República Federativa do Brasil*, 2020. Disponível em: <https://www.in.gov.br/web/dou/-/resolucao-n-77-de-22-de-setembro-de-2020-279178233>. Citado na página 18.

BRASIL, C. de Seguros Aliança do. *SEGURO AGRÍCOLA SEM COBERTURA DO FERS - Condições Contratuais - Versão 1.5*. 2022. Citado na página 16.

BROCKMAN, G. et al. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016. Citado na página 67.

CHACON, S.; STRAUB, B. *Pro git*. [S.l.]: Apress, 2014. Citado na página 67.

CHAN, S. C. et al. Measuring the reliability of reinforcement learning algorithms. *arXiv preprint arXiv:1912.05663*, 2019. Citado na página 43.

CHOI, J.-M.; KIM, J.-H.; KIM, S.-J. Application of reinforcement learning in detecting fraudulent insurance claims. *International Journal of Computer Science & Network Security*, International Journal of Computer Science & Network Security, v. 21, n. 9, p. 125–131, 2021. Citado 2 vezes nas páginas 26 e 33.

DICKS, D. L.; GARVEN, J. R.; HILLIARD, J. I. Optimism bias and the demand for insurance. *Available at SSRN 1907463*, 2018. Citado na página 25.

ERICSON, K. M.; SYDNOR, J. The questionable value of having a choice of levels of health insurance coverage. *Journal of Economic Perspectives*, v. 31, n. 4, p. 51–72, 2017. Citado na página 25.

GARBARINO, E.; SLONIM, R.; VILLEVAL, M. C. Loss aversion and lying behavior. *Journal of Economic Behavior Organization*, v. 158, p. 379–393, 2019. ISSN 0167-2681. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016726811830338X>. Citado na página 25.

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 67.

KAHNEMAN, D.; KNETSCH, J. L.; THALER, R. H. The endowment effect, loss aversion, and status quo bias: Anomalies. *Journal of Economic Perspectives*, v. 5, n. 1, p. 193–206, 1991. Citado na página 25.

KAHNEMAN, D.; LOVALLO, D. A cognitive perspective on risk taking. *Management Science*, Institute of Management Sciences, v. 39, n. 1, p. 17–31, 1993. Citado na página 25.

KAHNEMAN, D. et al. Sounding the alarm on system noise. *The McKinsey Quarterly*, McKinsey & Company, Inc., 2021. Citado 2 vezes nas páginas 18 e 26.

KAIRIES-SCHWARZ, N. et al. Health insurance choice and risk preferences under cumulative prospect theory–an experiment. *Journal of Economic Behavior & Organization*, Elsevier, v. 137, p. 374–397, 2017. Citado na página 25.

KRASHENINNIKOVA, E. et al. Reinforcement learning for pricing strategy optimization in the insurance industry. *Engineering applications of artificial intelligence*, Elsevier, v. 80, p. 8–19, 2019. Citado 5 vezes nas páginas 25, 28, 31, 53 e 68.

LU, F. Case study of an adaptive automated health insurance fraud auditor. *TRY ACAMEDICS!*, p. 52, 2010. Citado na página 26.

LU, F.; BORITZ, J. E.; COVVEY, D. Adaptive fraud detection using benford's law. In: SPRINGER. *Conference of the Canadian Society for Computational Studies of Intelligence*. [S.l.], 2006. p. 347–358. Citado 3 vezes nas páginas 32, 45 e 69.

MAPA. *REQUISITOS BÁSICOS PARA CAPACITAÇÃO DE PERITOS RURAIS*. 2020. Ministério da Agricultura, Pecuária e Abastecimento. Disponível em: <https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/seguro-rural/seguro-rural>. Citado na página 24.

_____. *Programas de apoio ao Seguro Rural*. Brasília, DF, 2021. Ministério da Agricultura, Pecuária e Abastecimento. Secretaria de Política Agrícola. Departamento de Gestão de Riscos. Disponível em: <https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/seguro-rural/publicacoes-seguro-rural/apresentacao-programas-de-apoio-ao-seguro-rural>. Citado 4 vezes nas páginas 15, 16, 67 e 71.

_____. *Guia de Seguros Rurais*. Brasília, DF, 2022. Ministério da Agricultura, Pecuária e Abastecimento. Disponível em: <https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/seguro-rural/publicacoes-seguro-rural/guia-do-seguro-rural-2022>. Citado 4 vezes nas páginas 12, 16, 18 e 19.

_____. *Histórico de Perdas na Agricultura Brasileira*. Brasília, DF, 2022. Ministério da Agricultura, Pecuária e Abastecimento. Secretaria de Política Agrícola. Departamento de Gestão de Riscos. Disponível em: <https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/seguro-rural/publicacoes-seguro-rural/historico-de-perdas-na-agricultura-brasileira-2000-2021.pdf>. Citado na página 20.

MCKINNEY, W. et al. Data structures for statistical computing in python. In: AUSTIN, TX. *Proceedings of the 9th Python in Science Conference*. [S.l.], 2010. v. 445, p. 51–56. Citado na página 67.

MEAD, A. et al. Detecting fraud in adversarial environments: A reinforcement learning approach. In: IEEE. *2018 Systems and Information Engineering Design Symposium (SIEDS)*. [S.l.], 2018. p. 118–122. Citado 5 vezes nas páginas 26, 29, 32, 53 e 69.

MILLER, S. J. et al. A quick introduction to benford's law. *Benford's Law: Theory and Applications*, Princeton University Press: Princeton, p. 3–22, 2015. Citado 2 vezes nas páginas 38 e 60.

MNIH, V. et al. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. Citado na página 61.

_____. Human-level control through deep reinforcement learning. *nature*, Nature Publishing Group, v. 518, n. 7540, p. 529–533, 2015. Citado na página 60.

MOSAVI, A. et al. Comprehensive review of deep reinforcement learning methods and applications in economics. *Mathematics*, MDPI, v. 8, n. 10, p. 1640, 2020. Citado 2 vezes nas páginas 26 e 29.

POOLE, D. L.; MACKWORTH, A. K. *Artificial Intelligence: foundations of computational agents*. [S.l.]: Cambridge University Press, 2010. Citado na página 41.

ROSSUM, G. V.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697. Citado na página 67.

SANTOS, G. R. d.; FREITAS, R. E. Gasto público com a agricultura no brasil: uma abordagem a partir de dados agregados. Instituo de Pesquisa Econômica Aplicada (Ipea), 2017. Citado na página 15.

SEGURADORA, T. M. *SEGURO AGRÍCOLA - CUSTEIO (SEM COBERTURA DO FUNDO DE ESTABILIDADE DO SEGURO RURAL - FESR) CONDIÇÕES GERAIS*. 2019. Citado na página 16.

SINAYOBYE, J. O.; KIWANUKA, F.; KYANDA, S. K. A state-of-the-art review of machine learning techniques for fraud detection research. In: IEEE. *2018 IEEE/ACM symposium on software engineering in africa (SEiA)*. [S.l.], 2018. p. 11–19. Citado 4 vezes nas páginas 18, 26, 28 e 29.

Sistema FAEP. *Podcast #77 Os Caminhos do Seguro Rural*. 2023. Podcast. Disponível em: <https://www.youtube.com/watch?v=drX7aFJAOCo>. Citado na página 59.

SUPERINTENDENTE DA SUPERINTENDÊNCIA DE SEGUROS PRIVADOS - SUSEP. *CIRCULAR SUSEP No 251, de 15 de abril de 2004*. 2004. <https://www2.susep.gov.br/textos/circ251.htm>. Consultado em: 31/07/2023. Citado na página 17.

SUTTON, R. S.; BARTO, A. G. *Reinforcement learning: An introduction*. [S.l.]: MIT press, 2018. Citado 2 vezes nas páginas 7 e 30.

TABOSA, F. J. S.; FILHO, J. E. R. V. Análise espacial do programa de subvenção ao prêmio do seguro rural (psr) e seu impacto na área cultivada e na produtividade agrícola no brasil. Revista Econômica do Nordeste, 2021. Citado na página 20.

TVERSKY, A.; KAHNEMAN, D. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, Springer, v. 5, n. 4, p. 297–323, 1992. Citado na página 25.

Apêndices

APÊNDICE A – Dados Originais

Tabela 17 – Grupos de Variáveis

Grupo	Variáveis	Descrição
Informações de Proposta e Apólice	*dt_proposta*, *id_proposta*	Data e identificação da proposta de seguro.
	dt_apolice, *ano_apolice*	Data e ano da apólice de seguro.
	nr_apolice, *cd_processo_susep*	Número da apólice e código do processo SUSEP.
	dt_inicio_vigencia, *dt_fim_vigencia*	Datas de início e fim da vigência da apólice.
	nr_proposta	Número da proposta de seguro.
Informações de Localização	*nr_min_long*, *nm_municipio_propriedade*	Longitude mínima e nome do município da propriedade.
	nr_seg_long, *nr_seg_lat*	Longitude e latitude da propriedade segurada.
	nr_grau_long, *nr_min_lat*	Grau de longitude e latitude mínima.
	nr_grau_lat, *latitude*	Grau de latitude e latitude da propriedade.
	longitude, *sg_uf_propriedade*	Longitude da propriedade e sigla do estado da propriedade.
	cd_geocmu, *nr_decimal_latitude*	Código geográfico do município e latitude decimal.
	nr_decimal_longitude, *formal_latitude*	Longitude decimal e latitude formal.
	formal_longitude	Longitude formal.
Informações de Segurado	*nm_segurado*, *nr_documento_segurado*	Nome e número do documento do segurado.
	nm_razao_social	Razão social do segurado.
Informações de Produto e Cultura	*nm_cultura_global*, *nm_classif_produto*	Nome da cultura global e classificação do produto segurado.
	nr_produtividade_segurada, *nr_produtividade_estimada*	Produtividade segurada e estimada do produto.
Informações de Cobertura e Prêmio	*niveldecobertura*, *pe_taxa*	Nível de cobertura do seguro e taxa de prêmio.
	vl_premio_liquido, *vl_limite_garantia*	Valor do prêmio líquido e limite de garantia.
	vl_subvencao_federal	Valor da subvenção federal.
Informações de Evento e Indenização	*evento_preponderante*, *valor_indenizacao*	Evento preponderante e valor da indenização.
Outras Informações	*nr_area_total*, *nr_animal*	Área total segurada e número de animais.

APÊNDICE B – Métricas de Performance Agente DQN por episódio

Tabela 18 – Resultados Experimentais: DQN vs. Aleatório

Experimento	**Inicial**	**Episódios**	**Acurácia**	**Precisão**	**Revocação**	**F1**
0	1 M	1	0.39	0.38	0.54	0.44
1	1 M	2	0.44	0.35	0.23	0.27
2	1 M	3	0.45	0.39	0.26	0.31
3	1 M	4	0.29	0.26	0.11	0.16
4	1 M	5	0.25	0.40	0.10	0.16
5	1 M	6	0.36	0.36	0.06	0.10
6	1 M	7	0.39	0.50	0.04	0.08
7	1 M	8	0.48	0.48	0.04	0.07
8	1 M	9	0.27	0.39	0.02	0.04
9	1 M	10	0.37	0.34	0.01	0.02
10	5 M	1	0.46	0.39	0.72	0.51
11	5 M	2	0.44	0.40	0.04	0.07
12	5 M	3	0.44	0.41	0.83	0.55
13	5 M	4	0.52	0.44	0.49	0.46
14	5 M	5	0.48	0.48	0.47	0.48
15	5 M	6	0.49	0.48	0.44	0.46
16	5 M	7	0.44	0.42	0.39	0.40
17	5 M	8	0.35	0.35	0.24	0.29
18	5 M	9	0.49	0.42	0.35	0.38
19	5 M	10	0.46	0.42	0.12	0.18
20	10 M	1	0.46	0.37	0.21	0.27
21	10 M	2	0.45	0.41	0.30	0.35
22	10 M	3	0.39	0.40	0.36	0.38
23	10 M	4	0.48	0.45	0.30	0.36
24	10 M	5	0.46	0.38	0.47	0.42
25	10 M	6	0.38	0.41	0.28	0.33
26	10 M	7	0.22	0.38	0.03	0.06
27	10 M	8	0.48	0.44	0.43	0.44
28	10 M	9	0.47	0.43	0.38	0.40
29	10 M	10	0.42	0.41	0.44	0.43
30	15 M	1	0.38	0.44	0.11	0.18
31	15 M	2	0.41	0.40	0.31	0.35
32	15 M	3	0.45	0.42	0.21	0.28
33	15 M	4	0.53	0.49	0.60	0.54
34	15 M	5	0.42	0.40	0.35	0.37
35	15 M	6	0.31	0.44	0.05	0.10
36	15 M	7	0.48	0.44	0.43	0.44
37	15 M	8	0.49	0.43	0.48	0.46
38	15 M	9	0.42	0.40	0.05	0.08
39	15 M	10	0.42	0.44	0.53	0.48