UNIVERSIDADE FEDERAL DA PARAÍBA (UFPB)
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS (CCSA)
DEPARTAMENTO DE FINANÇAS E CONTABILIDADE (DFC)
CURSO DE BACHARELADO EM CIÊNCIAS ATUARIAIS (CCA)

CRISTIANE SIQUEIRA DA SILVA

COMPARAÇÃO ENTRE MODELOS DE REGRESSÃO ESTATÍSTICOS E MODELOS DE APRENDIZADO DE MÁQUINA INTERPRETÁVEIS NA MODELAGEM DE RISCO NO SEGURO AGROPECUÁRIO

JOÃO PESSOA - PB
2026

CRISTIANE SIQUEIRA DA SILVA

COMPARAÇÃO ENTRE MODELOS DE REGRESSÃO ESTATÍSTICOS E MODELOS DE APRENDIZADO DE MÁQUINA INTERPRETÁVEIS NA MODELAGEM DE RISCO NO SEGURO AGROPECUÁRIO

Trabalho de Conclusão de Curso para o curso de Ciências Atuariais do Centro de Ciências Sociais Aplicadas, da Universidade Federal da Paraíba.

Orientador: Prof.º Dr. Filipe Coelho de Lima Duarte.

JOÃO PESSOA - PB

2026

Catalogação na publicação
Seção de Catalogação e Classificação

S586c Silva, Cristiane Siqueira da.
Comparação entre modelos de regressão estatísticos e modelos de aprendizado de máquina interpretáveis na modelagem de risco no seguro agropecuário / Cristiane Siqueira da Silva. - João Pessoa, 2026.
79 f.

Orientação: Filipe Coelho de Lima Duarte.
TCC (Graduação) - UFPB/CCSA.

1. Seguro rural. 2. Severidade de sinistros. 3. Aprendizado de máquina. 4. Modelos lineares generalizados. 5. Explicabilidade (SHAP). I. Duarte, Filipe Coelho de Lima. II. Título.

UFPB/CCSA CDU 368(043)

Elaborado por ANDRE DOMINGOS DA SILVA FERNANDES - CRB-15/00730

CRISTIANE SIQUEIRA DA SILVA

COMPARAÇÃO ENTRE MODELOS DE REGRESSÃO ESTATÍSTICOS E MODELOS DE APRENDIZADO DE MÁQUINA INTERPRETÁVEIS NA MODELAGEM DE RISCO NO SEGURO AGROPECUÁRIO

Trabalho de Conclusão de Curso para o curso de Ciências Atuariais do Centro de Ciências Sociais Aplicadas, da Universidade Federal da Paraíba.

BANCA EXAMINADORA

Documento assinado digitalmente
gov.br
FILIPE COELHO DE LIMA DUARTE
Data: 10/08/2026 20:55:03-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Filipe Coelho de Lima Duarte

Orientador

Universidade Federal da Paraíba (UFPB)

Documento assinado digitalmente
gov.br
YURI MARTI SANTANA SANTOS
Data: 10/08/2026 12:39:39-0300
Verifique em https://validar.iti.gov.br

Prof. Me. Yuri Martí Santana Santos

Membro Avaliador

Universidade Federal da Paraíba (UFPB)

Documento assinado digitalmente
ICP Brasil
SAMARA LAUAR SANTOS
Data: 10/08/2026 10:22:26-0300
Verifique em https://validar.iti.gov.br

Profa. Ma. Samara Lauar Santos

Membro Avaliador

Universidade Federal da Paraíba (UFPB)

AGRADECIMENTOS

Primeiramente, agradeço ao meu bom Deus, por ter me dado força para enfrentar todas as adversidades que me fizeram pensar em desistir.

Agradeço à minha mãe, Elione Alves, que apesar de tudo sempre esteve do meu lado, me dando forças e palavras de conforto e consolação. Ao meu pai, João Roberto (*in memoriam*), que apesar de nunca o conhecer, sempre ouvi o quão bom era e como era querido pelas pessoas, e fez-me sentir muito orgulhosa de ser sua filha.

Agradeço aos meus professores da escola Carlos Chagas que me fizeram enxergar o quão importante a educação é na vida das pessoas, por me fazer gostar de estudar e sempre terem me ajudado.

Agradeço a todas as pessoas que conheci ao longo da minha vida, em especial aos meus queridos amigos, Carlos, Cléo, Margarett, que me ajudaram muito e não tenho palavras para agradecer a ajuda, a força, os conselhos, amo muito vocês.

Agradeço, por fim, aos meus professores da graduação, Luiz, Vera, Filipe, que me fizeram gostar cada vez mais do curso, por abrir a minha mente para novas possibilidades, pela força, pelos conselhos e pelas oportunidades.

Muito obrigado a todos.

RESUMO

O seguro agrícola constitui instrumento relevante de mitigação do risco climático e produtivo no setor agropecuário brasileiro, sendo a modelagem da severidade dos sinistros elemento central para a precificação e a gestão de risco das seguradoras. Este trabalho teve como objetivo comparar modelos de regressão e modelos de aprendizado de máquina interpretáveis na previsão da severidade de sinistros do seguro rural, avaliando tanto o desempenho preditivo quanto a interpretabilidade das abordagens. Utilizou-se uma base de 190.423 apólices com ocorrência de sinistro, extraída do Sistema de Subvenção Econômica ao Prêmio do Seguro Rural (SISSER), referente ao período de 2016 a 2024. Foram estimados quatro modelos: Regressão Linear Múltipla (RLM), Modelo Linear Generalizado (MLG) com distribuição Gama, Random Forest e XGBoost, avaliados por validação cruzada e comparados pelas métricas MAE, RMSE e sMAPE. A interpretabilidade dos modelos de aprendizado de máquina foi analisada por meio da técnica SHAP (SHapley Additive exPlanations). Os resultados mostraram que os modelos de aprendizado de máquina superaram os estatísticos em desempenho preditivo, com redução de aproximadamente 12% no RMSE em relação ao MLG Gama, ainda que com maior diferença entre os resultados de treino e teste; entre eles, o XGBoost obteve os menores valores de MAE e RMSE, e o Random Forest, o menor sMAPE. O limite de garantia e o ano de 2021, marcado por severa estiagem, destacaram-se como os principais determinantes da severidade em todos os modelos, resultado convergente entre as abordagens estatística e de aprendizado de máquina. A segmentação da base entre atividades agrícola e pecuária não produziu ganho de desempenho, dado o predomínio das apólices agrícolas na amostra. Concluiu-se que a combinação de modelos de aprendizado de máquina com técnicas de explicabilidade pode constituir uma alternativa promissora para conciliar precisão preditiva e transparência na modelagem do risco agropecuário.

Palavras-chave: Seguro Rural. Severidade de Sinistros. Aprendizado de Máquina. Modelos Lineares Generalizados. Explicabilidade (SHAP).

ABSTRACT

Agricultural insurance is a relevant instrument for mitigating climatic and production risk in the Brazilian agribusiness sector, with claim severity modeling being a central element for insurers' pricing and risk management. This study aimed to compare regression models and interpretable machine learning models in predicting the severity of rural insurance claims, evaluating both predictive performance and the interpretability of the approaches. A database of 190,423 policies with claim occurrence was used, drawn from the Rural Insurance Premium Subsidy System (SISSER), covering the period from 2016 to 2024. Four models were estimated: Multiple Linear Regression (MLR), Generalized Linear Model (GLM) with Gamma distribution, Random Forest, and XGBoost, evaluated through cross-validation and compared using the MAE, RMSE, and sMAPE metrics. The interpretability of the machine learning models was assessed using the SHAP (SHapley Additive exPlanations) technique. Results showed that the machine learning models outperformed the statistical ones in predictive performance, with an approximate 12% reduction in RMSE compared to the Gamma GLM, albeit with greater relative overfitting between training and testing; among them, XGBoost achieved the lowest MAE and RMSE, and Random Forest the lowest sMAPE. The guarantee limit and the year 2021, marked by severe drought, stood out as the main determinants of severity across all models, a result that converged between the statistical and machine learning approaches. Segmenting the database between agricultural and livestock activities produced no performance gain, given the predominance of agricultural policies in the sample. It was concluded that combining machine learning models with explainability techniques constitutes a viable alternative for reconciling predictive accuracy and transparency in agricultural risk modeling.

Keywords: Rural Insurance. Claim Severity. Machine Learning. Generalized Linear Models. Explainability (SHAP).

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SUSEP - Superintendência de Seguros Privados

CNSP - Conselho Nacional de Seguros Privados

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária

MLG - Modelos Lineares Generalizados

AM – Aprendizado de máquina

SHAP - Explicações Aditivas de Shapley

RLM – Regressão Linear Multipla

MAE - Erro Absoluto Médio

RMSE - Raiz do Erro Quadrático Médio

sMAPE – Erro Percentual Absoluto Médio Simétrico

PIB - Produto Interno Bruto

ZARC - Zoneamento Agrícola de Risco Climático

MAPA - Ministério da Agricultura, Pecuária e Abastecimento

ARIMA - Modelo Autorregressivo Integrado de Médias Móveis

RF – *Random Forest*

XGBoost - *Boosting* de Gradiente Extremo

LIME - Explicações Locais Interpretáveis Independentes de Modelo

SISSER - Sistema de Subvenção Econômica ao Prêmio do Seguro Rural

PSR - Programa de Subvenção ao Prêmio do Seguro Rural

IPCA - Índice Nacional de Preços ao Consumidor Amplo

SIDRA - Sistema IBGE de Recuperação Automática

FIV - Fator de inflação da variância

SUMÁRIO

1 INTRODUÇÃO

Conforme dispõe o art. 8 da Circular SUSEP n° 682/2022, o Seguro Rural é considerado um instrumento importante da política agrícola nacional, permitindo o produtor rural proteger-se contra prejuízos causados por fenômenos climáticos adversos. Seu objetivo principal é oferecer cobertura adequada às distintas necessidades do campo, contribuindo para a mitigação do risco. A Resolução do Conselho Nacional de Seguros Privados (CNSP) n.º 404/2021 define oito modalidades de Seguro Rural, sendo as mais empregadas no Brasil o Seguro Agrícola e o Seguro Pecuário (Caffagni, Paixão e Rios, 2022).

A agropecuária, entendida como o estudo, teoria e aplicação da agricultura e pecuária em suas interações mútuas, é inserida no setor primário da economia, exercendo função de importância na economia internacional e brasileira, configurando-se o setor de maior crescimento. Segundo o Ministério da Agricultura e Pecuária, a agropecuária brasileira alcançou um valor de produção de R$ 1,189 trilhões em 2022, sendo R$ 814,77 bilhões da agricultura e R$ 374,27 bilhões da pecuária, demonstrando a sua importância na economia nacional. No entanto, este setor é o mais vulnerável aos riscos climáticos, segundo a Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA, 2023), a agropecuária brasileira é afetada diretamente por fatores climáticos como chuva, umidade de solo e ar, ventos, temperatura e radiação solar.

O risco pode ser compreendido como a probabilidade de ocorrência de um evento adverso, mas também como uma possibilidade incerta, assumindo conotação quantitativa — probabilidade mensurável — e conotação qualitativa — incerteza e imprevisibilidade (Almeida, 2023). Na agropecuária, o risco relaciona-se a fatores climáticos; as alterações climáticas geram modificações na ocorrência e frequência de eventos extremos, como inundações, secas e geadas (Tanure *et al.*, 2024). Além disso, essas transformações afetam o ciclo de pragas e doenças nas culturas e nos animais, impactando diretamente a produção (EMBRAPA, 2023).

Neste contexto, a mitigação do risco no setor agropecuário é fundamental para o desenvolvimento econômico e produtivo do Brasil. Assim, como meio de mitigação, as seguradoras recorrem a modelos estatísticos tradicionais (Kratka, 2015). No entanto, com o avanço tecnológico e o grande volume de dados, se torna relevante a utilização de modelos preditivos para a mitigação e gerenciamento do risco por parte das seguradoras (Lucas *et al.*, 2025).

A modelagem estatística é utilizada no setor securitário para estimar a severidade dos sinistros, quantificar riscos e definir provisões técnicas, contribuindo para maior precisão na precificação e no gerenciamento das perdas (Navatha, 2022). São utilizados métodos estatísticos, como: Modelo de Regressão Linear, análise de Séries Temporais, Modelos de Regressão não Linear, Modelos Lineares Generalizados (MLG), entre outros, que se destacam pela interpretabilidade, confiabilidade e bom desempenho em base de dados pequenas (Steiner; Meng, 2019). Apesar disso, enfrentam limitações na captura de padrões complexos e não lineares em volumes grandes de dados e apresentam desempenho inferior em base de dados desbalanceados e com ruído (Aas *et al.*, 2024).

Atualmente, diversas áreas do conhecimento estão utilizando o aprendizado de máquina para resolver problemas de previsão e classificação, pois são capazes de lidar com grandes volumes de dados e aprender padrões complexos sem programação explícita (Rifaldi, 2025; Vinuesa *et al.*, 2025). Esta capacidade permite prever com precisão os custos do seguro para a apólice, otimizando o tempo e aumentando o rendimento das seguradoras de maneira eficiente (Shinde e Raut, 2018; Boehm; Kumar; Yang, 2019). No entanto, o desempenho do modelo pode ser prejudicado pelo *overfitting,* restringindo a capacidade de generalização para novos dados e prejudicando a robustez dos modelos (Ying, 2019).

No aprendizado de máquina (AM) existem algoritmos não interpretáveis que, apesar do ótimo desempenho analítico e preditivo, apresentam grande complexidade e baixa interpretabilidade (Hassija *et al.*, 2024). Diante dessa limitação, surgem os modelos de AM interpretáveis, que buscam conciliar precisão e interpretabilidade ao permitir que se compreenda como cada variável contribui para a previsão. Essa interpretabilidade, alcançada por meio de técnicas como SHAP e outras abordagens explicativas, auxilia os tomadores de decisão a entenderem o raciocínio do modelo, tornando sua utilização mais segura e adequada em aplicações práticas (Gao; Guan, 2023).

Nesse contexto, torna-se importante os métodos de modelagem e a necessidade de transparência nas decisões dos modelos AM. Diante disso, a questão central deste estudo é: **Como os modelos estatísticos tradicionais se comparam aos modelos de aprendizado de máquina interpretáveis na modelagem do risco do seguro agropecuário?**

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Comparar a eficácia do desempenho e interpretabilidade dos modelos de regressão e dos modelos interpretáveis de aprendizado de máquina na modelagem do risco agropecuário.

1.1.2 Objetivos Específicos

a) Implementar os modelos estatísticos e de aprendizado de máquina no processo de previsão da severidade de sinistros no seguro agropecuário;

b) Confrontar o desempenho de previsão dos modelos aplicados por meio das métricas de erro *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE) e *Symmetric Mean Absolute Percentage Error* (sMAPE);

c) Analisar a interpretabilidade dos modelos confrontando a clareza e a consistência das explicações fornecidas por cada modelo.

1.2 JUSTIFICATIVA

A agropecuária brasileira representa um dos pilares econômicos do país, respondendo por cerca de 25% do Produto Interno Bruto (PIB) nacional (IBGE, 2024). Além disso, é responsável por grandes volumes de exportações, com superávit de mais R$ 148 bilhões em 2023, consolidando o Brasil como um dos maiores produtores de *commodities* como soja, milho, carne bovina e café (MAPA, 2024). Essa expressividade torna o setor crucial para a segurança alimentar e para o equilíbrio comercial, além de estimular o desenvolvimento tecnológico no campo (EMBRAPA, 2024).

Segundo Tanure *et al.* (2024), embora a agropecuária seja economicamente robusta, encontra-se constantemente vulnerável aos efeitos adversos das mudanças climáticas e aos ciclos de pragas e doenças nas culturas e nos animais, impactando diretamente a produção (EMBRAPA, 2023). O autor ressalta que, entre os anos de 2023 e 2025, o Brasil sofreu um aumento de eventos climáticos, como secas prolongadas, ondas de calor e chuvas torrenciais, tais eventos impactam fortemente a produtividade de culturas e criação de gado. Conforme o

Instituto Clima e Sociedade (2025), o Brasil acumulou quase R$ 300 bilhões em perdas entre 2013 e 2022 devido a eventos climáticos extremos. Portanto, torna-se indispensável a escolha de modelos que possam ajudar as seguradoras na mitigação dos riscos desse setor (Zhichkin *et al.*, 2023).

Os modelos estatísticos tradicionais ainda prevalecem entre as seguradoras que atuam no setor agropecuário (Navatha, 2022). No entanto, esses modelos enfrentam limitações consideráveis diante da complexidade dos riscos climáticos, especialmente pela dificuldade em capturar padrões não lineares – tornando necessário construir a fórmula que captura este tipo de relação, no qual depende de conhecimento prévio –, interdependência espacial e efeitos extremos (Aas *et al.*, 2024). Essas limitações se tornam críticas em cenários de eventos climáticos intensos, como secas ou enchentes simultâneas, que prejudicam várias regiões do país (Zhichkin, 2023). Em contrapartida, a essas limitações, estudos recentes vêm trazendo abordagens mais robustas, como modelos bayesianos incorporados ao Zoneamento Agrícola de Risco Climático (ZARC) (Pereira *et al.*, 2025) e técnicas baseadas em aprendizado de máquina interpretáveis (Swain *et al.*, 2025).

A adoção de modelos de aprendizado de máquina no setor securitário vem avançando continuamente, porém a utilização prática desses modelos ainda enfrenta barreiras quando os resultados não são facilmente compreensíveis pelos analistas e gestores (Pedulla, 2022). Nesse contexto, as técnicas de interpretabilidade utilizadas em modelos não interpretáveis torna-se um fator estratégico para a tomada de decisão, pois auxilia na compreensão de como as variáveis influenciam na previsão do risco possibilitando decisões mais seguras, rastreáveis e alinhadas às exigências regulatórias do setor (Swain *et al.*, 2025; Zhichkin *et al.*, 2023). Além disso, as técnicas de interpretabilidade contribuem em ajustes mais precisos e na redução de erros associados ao treinamento dos modelos, como o *overfitting* (Cartolano *et al.*, 2024; Biaugini, 2023).

Portanto, o presente trabalho apresenta qual modelo demonstra maior eficácia na modelagem dos riscos agropecuários, contribuindo assim, para a melhoria da mitigação do risco no setor segurador, que consequentemente, amplia a precisão da precificação, e a compreensão das variáveis envolvidas para o melhor desempenho das seguradoras no mercado.

2 REVISÃO DA LITERATURA

2.1 SEGURO AGROPECUÁRIO

No Brasil, a agricultura e a pecuária exercem um papel central na economia, sendo essenciais na balança comercial e no aumento do Produto Interno Bruto (PIB) nacional, representando cerca de 25% do PIB (IBGE, 2024). Em 2025, a estimativa é que a agropecuária represente cerca de 29,4% do PIB, destacando-se que no primeiro trimestre de 2025 o ramo agrícola cresceu em 5,59% o PIB nacional e o ramo da pecuária em 8,50% (CEPEA, 2025).

O setor agrícola e pecuário movimentou em 2024 cerca de US$ 164,4 bilhões, este montante corresponde a 49% das exportações totais brasileiras, destacando as *commodities:* soja, milho, café, carne suína, carne bovina, entre outros (MAPA, 2025). Além disso, a agropecuária gera milhares de empregos diretos e indiretos (FGV, 2025), além de contribuir para o fortalecimento das cadeias produtivas e inovação tecnológica no campo (Barbosa; Brisola, 2024).

No relatório "VISÃO 2030: O Futuro da Agricultura Brasileira" elaborado pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), em conjunto com o Ministério da Agricultura, Pecuária e Abastecimento (MAPA), distingue a agropecuária de outros setores da economia brasileira, pois depende primordialmente de recursos naturais e processos biológicos, destacando as mudanças climáticas ao longo dos anos e seu impacto negativo no solo fértil (EMBRAPA; MAPA, 2018).

O relatório retrata que 75% dos alimentos no mundo são produzidos a partir de 12 espécies de plantas e cinco espécies de animais, tornando o sistema alimentar global extremamente vulnerável aos riscos da atividade agrícola, como pragas e doenças em animais e plantas, agravados pelas mudanças climáticas. Os eventos climáticos podem resultar em perdas significativas na produção, queda das exportações, redução da ocupação direta e indireta, maior volatilidade na produção e renda dos produtores, aumento dos preços para os consumidores e insegurança alimentar (FAO; OPAS; CEPAL., 2025).

O autor Tanure *et al.* (2024) ressalta igualmente a vulnerabilidade da agropecuária aos efeitos das mudanças climáticas, contextualizando que entre os anos de 2023 e 2025, o Brasil sofreu com eventos climáticos extremos, impactando os cultivos e animais. Além disso, o Instituto de Clima e Sociedade (2025) evidencia a perda acumulada do Brasil em R$ 300 bilhões entre os anos de 2013 e 2022 devido aos eventos climáticos extremos.

Em decorrência disso, o seguro é considerado uma ferramenta para evitar o máximo de perdas possíveis em diversas áreas. Neste contexto, o seguro pode ser compreendido como um contrato no qual a seguradora assume o pagamento da indenização ao segurado caso ocorra um evento previamente estabelecido, para isto, o segurado paga o prêmio a seguradora, ou seja, a seguradora é responsável por emitir a apólice e assumir o risco (BRASIL, 2024). No Brasil, existem 13 ramos e seguros, destacam-se os seguros de automóvel, responsabilidade, pessoas, danos, transporte e rural (SUSEP, 2024).

Neste contexto, o Seguro Rural é responsável por proteger o produtor rural dos riscos de eventos extremos causados pelos efeitos climáticos (Pereira, 2025). Segundo a resolução do CNSP n.º 404/2021 e a Circular SUSEP n° 682/2022, o seguro rural é constituído por oito modalidades: Seguro agrícola; Seguro pecuário; Seguro aquícola; Seguro de florestas; Seguro de penhor rural; Seguro de benfeitorias e produtos agropecuários e Seguro de vida do produtor rural, sendo a modalidade de Seguro Rural mais empregada no Brasil é o Seguro Agrícola e o Seguro Pecuário (Caffagni; Paixão; Rios, 2022).

O seguro agrícola protege a produção dos produtores rurais de eventos climáticos extremos como: geada, seca, granizo, tempestade, entre outros (Zhichkin *et al.*, 2023) e o seguro pecuário voltado a proteger os criadores de animais dos riscos referente à morte, doença, acidentes e roubo (Dal Pozzo *et al.*, 2023).

Consoante a isto, as seguradoras utilizam a modelagem de risco para auxiliá-las na tomada de decisão ao oferecer um contrato de seguro, levando em consideração riscos relacionados a diferentes situações financeiras ou operacionais (Pereira *et al.*, 2025).

2.2 MODELAGEM DE RISCO

Segundo Galvão *et al.* (2018), o risco é compreendido como uma métrica relativa a possíveis perdas impostas aos entes econômicos[1] diante das incertezas ocasionadas pelas suas atividades. Além disso, o autor define o gerenciamento ou gestão do risco como um compilado de pessoas, métricas de controle e sistemas conduzidos a avaliar e controlar os riscos associados ao ente econômico.

Consoante a isto, a modelagem de risco é definida como um processo analítico visando a identificação, quantificação e mitigação dos riscos relacionados a distintas situações financeiras e operacionais, sendo importante nas áreas de finanças, seguros e gestão de projetos,

[1] Denominados de entes econômicos por estarem expostos a danos ou perdas de valor econômico mensurável.

em que as incertezas impactam os resultados (Pereira *et al.*, 2025). A modelagem de risco pode utilizar técnicas de estatística, matemática e de aprendizado de máquina para prever a probabilidade, ocorrência e severidade de eventos adversos e suas consequências para as empresas, permitindo a tomada de decisão eficiente (Taero, 2016). A modelagem de risco é um fator crucial para as empresas, pois permite compreender os riscos envolvidos em suas operações e assim desenvolver estratégias para mitigá-lo permitindo a preparação para os cenários adversos aumentando a persistência em tempos de crise (Pereira *et al.*, 2025). No entanto, a modelagem enfrenta limitações, como acesso a dados de qualidade, modelos complexos que impedem a interpretação dos resultados, premissas do modelo e o evento do cisne negro, dificuldade de capturar eventos imprevistos, (Souza; Assunção, 2020).

A modelagem de risco pode ser organizada em quatro áreas principais. Os riscos de mercado referem-se à exposição às flutuações nos preços de ativos, nas taxas de juros, no câmbio e nas commodities (Galvão; Oliveira; Fleuriet, 2018). Já os riscos de crédito avaliam a probabilidade de inadimplência das contrapartes, ou seja, a chance de que uma instituição ou indivíduo não honre seus compromissos financeiros (Schmidt *et al.*, 2026). Os riscos sistêmicos dizem respeito à possibilidade de colapso de todo o sistema financeiro ou econômico, com impactos generalizados e interdependentes (Martins, 2020). Por fim, os riscos de subscrição envolvem a precificação de prêmios e a constituição de reservas técnicas para cobrir sinistros futuros, garantindo a sustentabilidade das operações (Shams; Ghosh, 2026).

No contexto do setor segurador, essa modelagem é compreendida como o processo de aplicação dos métodos e técnicas estatísticos, matemáticos e computacionais para identificar, quantificar e prever a probabilidade e o impacto de eventos que são capazes de gerar perdas financeiras para as seguradoras (Shams; Ghosh, 2026). Esse procedimento contribui na precificação de apólices, definição de reservas técnicas e na tomada de decisão estratégica reduzindo a incerteza e garantindo o equilíbrio financeiro da operação (Nagy; Cotlet, 2011).

Esse processo também envolve a análise de dados históricos sobre sinistro, prêmios, perfis de clientes, condições de risco, construção de modelos preditivos, além de simular cenários que possibilitem avaliar o comportamento do portfólio em diferentes condições (Shams; Ghosh, 2026). Neste contexto, a modelagem preditiva vem sendo amplamente utilizada por vários setores como *marketing*, saúde, finanças e seguro, pois utiliza métodos estatísticos e de aprendizado de máquina que ajudam a prever resultados futuros auxiliando na tomada de decisão (Burns *et al.*, 2019).

2.2.1 Modelagem preditiva

A modelagem preditiva é compreendida como o processo de construção de modelos estatísticos ou de aprendizado de máquina com o objetivo de prever eventos futuros ou desconhecidos com base em dados históricos. Os modelos capturam relações entre variáveis, permitindo a projeção de tendências, comportamentos ou eventos futuros (Verbeke; Dejaeger; Martens, 2021).

Max Kuhn e Kjell Johnson (2013) conceituam a modelagem preditiva como o desenvolvimento de uma ferramenta matemática capaz de gerar previsões precisas e que para alcançar tais previsões são necessárias etapas interdependentes, começando pela definição clara do objeto de previsão, seguido pela aquisição e preparação dos dados, escolha e treinamento do modelo, ajuste de parâmetros, avaliação de desempenho e pôr fim a aplicação do modelo em cenários reais.

Figura 1 – Ilustração das etapas da Modelagem Preditiva.

Fonte: Nology (2024).

A Figura 1 demonstra o ciclo de desenvolvimento da modelagem preditiva, em que cada etapa faz parte de um processo contínuo para criar, treinar, melhorar e implementar os modelos preditivos.

Neste contexto, a modelagem preditiva é um método comumente utilizado por diversos setores, como negócios, saúde, educação, finanças e seguros, apoiando a tomada de decisão com base em evidências quantitativas (Burns *et al.*, 2019).

Eileen Burns *et al.* (2019) no relatório técnico da *Society of Actuaries*, destaca a modelagem preditiva como ferramenta estratégica no setor de seguros. Neste contexto, a

modelagem preditiva no setor de seguros utiliza uma abordagem analítica com dados históricos com o objetivo de prever o cancelamento de apólices, fraudes de clientes, padrões de pagamento e prever a frequência e severidade dos sinistros. Ressaltam a importância de prever a severidade[2] do sinistro no setor securitário, assim as seguradoras podem oferecer prêmios justos, adequar as reservas técnicas e melhorar estratégias de prevenção, além disso, destaca a utilização da modelagem estatística - Modelos Lineares Generalizados – e a modelagem de aprendizado de máquina – *Random Forest*, XGBoost – para prever a severidade do sinistro.

2.3 MODELAGEM ESTATÍSTICA

Segundo o autor Morettin e Bussab (2010), a Estatística pode ser compreendida como a ciência voltada para a organização, descrição, análise e interpretação dos dados experimentais, objetivando a tomada de decisão. Esta ciência é utilizada em diversas áreas do conhecimento que utilizam dados experimentais, servindo como base para entender fenômenos, identificar padrões e realizar previsões.

O autor Morettin (2023) ressalta que a estatística utiliza modelos como ferramenta para representar de forma matemática a relação entre variáveis, permitindo explicar comportamentos entre variáveis, realizar testes de hipótese e previsões e que através desses modelos, estruturados a partir de observações, interpretação e pressupostos sobre os dados, as empresas conseguem extrair o máximo de informações possíveis. Os principais modelos são: Modelos de Regressão, Séries Temporais, Classificação, Agrupamento, Probabilísticos, Bayesianos e Multivariados (Steiner; Meng, 2019).

Nesse contexto, destaca-se a modelagem estatística, entendida como o processo de construção, avaliação e seleção de modelos estatísticos adequados para representar um determinado fenômeno. Esse processo possibilita identificar padrões nos dados, antecipar tendências e subsidiar a tomada de decisões fundamentadas em evidências quantitativas (Morettin, 2023). Destacam-se as modelagens paramétricas, não paramétricas, semi paramétricas, determinística e estocástica (Becker, 2015).

A modelagem estatística é amplamente utilizada em diversos setores, inclusive no setor securitário, destacando-se as áreas de análise atuarial, precificação de seguros, cálculos das provisões técnicas e avaliação de risco (Navatha, 2022). No setor agropecuário, especificamente, a modelagem pode ser aplicada para prever safras, estimar produtividade,

[2] Valor financeiro do sinistro.

monitorar variações climáticas, entre outros, consolidando-se como ferramenta essencial no setor agropecuário (Liakos *et al.*, 2018).

Para Becker (2015), a modelagem estatística é indispensável para transformar dados em conhecimento aplicável. O autor destaca a relevância da modelagem na tomada de decisão de vários setores, ao transformar dados em previsões úteis, além de detalhar as limitações da modelagem que pode sofrer de dependência da qualidade dos dados, risco de sobreajuste (*overfitting*) e dificuldade de interpretação de modelos complexos.

Consoante a isto, os modelos regressão linear e os modelos lineares generalizados (MLG), são amplamente utilizados, podendo ser empregados para estimar a frequência e severidade de sinistros através das características dos segurados, além de serem modelos flexíveis e terem a capacidade de lidar com diferentes distribuições (Clemente *et al.*, 2023).

2.3.1 Modelos estatísticos Regressão linear e Modelos Lineares Generalizados

A regressão linear é considerada um dos métodos estatísticos mais tradicionais, podendo ser subdividida em simples, múltiplas, polinomial e com interações, e é utilizada para analisar a relação entre uma variável resposta contínua e um ou mais variáveis preditoras (explicativa), baseado na suposição de que a variável dependente segue uma distribuição normal e a relação com as variáveis explicativas pode ser expressa por uma combinação linear permitindo estimar como as mudanças nos preditores impactam a variável de interesse (Frees, 2010).

A regressão linear múltipla (RLM) é uma extensão do modelo de regressão linear simples, que utiliza múltiplas variáveis explicativas simultaneamente com o intuito de determinar a variável resposta, possibilitando avaliar o efeito conjunto de diferentes fatores sobre a variável resposta e controlar possíveis confusões entre preditores (Frees, 2010).

A regressão linear múltipla é demonstrada através da equação (1), e descreve a variável dependente sendo determinada por múltiplas variáveis explicativas (Gujarati; Porter, 2011).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \quad (1)$$

Onde:

Y_i : é o valor da variável resposta na i-ésima observação;

$\beta_0, \ldots, \beta_p$: são os parâmetros;

$X_{i1}, \ldots, X_{ip}$: são os valores das variáveis preditoras na i-ésima observação;

ε_i : é um termo de erro aleatório com distribuição normal, média zero e variância constante.

No entanto, apesar da regressão linear múltipla ser simples e fácil de interpretar, apresenta limitações importantes, pois exige homoscedasticidade, variância constante dos erros, normalidade dos resíduos e independência entre observações, quando essas limitações não são atendidas a qualidade das estimativas pode ser comprometida, apesar disso, é útil por sua clareza, baixo custo computacional e por fornecer resultados facilmente interpretáveis (Montgomery; Peck; Vining, 2012).

Quanto à violação dessas hipóteses, a literatura destaca que a heterocedasticidade não enviesa os coeficientes estimados, mas invalida os erros-padrão e, consequentemente, os testes de significância e os intervalos de confiança, tornando-os inconsistentes (Wooldridge, 2016). Os testes de Breusch-Pagan e de White são os procedimentos mais utilizados para diagnosticar heterocedasticidade, avaliando se a variância dos resíduos depende das variáveis explicativas (Breusch; Pagan, 1979; White, 1980); já a normalidade dos resíduos é usualmente verificada pelo teste de Jarque-Bera e por gráficos quantil-quantil (Q-Q plot), sendo sua violação menos crítica em amostras grandes, dado o Teorema Central do Limite, mas ainda relevante para a validade de inferências pontuais (Montgomery; Peck; Vining, 2012).

Consoante a isto, um meio de atenuar as limitações estão na utilização da regressão linear múltipla com logaritmo natural na variável dependente, essa transformação é extremamente útil quando a variável resposta assume apenas valores positivos e apresenta distribuição assimétrica, como ocorre em dados financeiros ou de custos (Klugman *et al.*, 2019). A utilização do logaritmo reduz a assimetria, estabiliza a variância e melhora o ajuste do modelo, além de permitir interpretações consistentes dos coeficientes em termos de variações proporcionais (Benoit *et al.*, 2011). De forma ampla, as transformações de variáveis constituem recurso metodológico relevante, englobando não apenas o logaritmo, mas também alternativas como raiz quadrada, padronização e Box-Cox, todas voltadas a melhorar a adequação dos dados às premissas dos modelos estatísticos (Gujarati; Porter, 2011).

A regressão linear múltipla com logaritmo natural na variável dependente é demonstrada através da equação (2), e descreve a variável dependente transformada sendo determinada por múltiplas variáveis explicativas (Gujarati; Porter, 2011).

$$ln\,(\,Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \qquad (2)$$

Onde:

$ln\ (Y_i)$: é o logaritmo natural de Y_i;

Y_i : é o valor da variável resposta na i-ésima observação;

$\beta_0, \ldots, \beta_p$: são os parâmetros;

$X_{i1}, \ldots, X_{ip}$: são os valores das variáveis preditoras na i-ésima observação;

ε_i : é um termo de erro aleatório com distribuição normal, média zero e variância constante.

No setor securitário esta abordagem aplica-se a previsão da severidade dos sinistros, ou seja, a previsão do valor das indenizações, como os custos dos sinistros geralmente são concentrados em valores baixos, mas com ocorrência de indenizações muito altas, a transformação logarítmica torna as estimativas mais próximas da distribuição real dos dados (Klugman *et al.*, 2019). Embora a regressão linear múltipla se destaque por ter interpretação direta e menor custo computacional e se destaque em aplicações práticas, é uma técnica simplista em comparação aos modelos lineares generalizados (Blei, 2015).

Deste modo, os modelos lineares generalizados (MLG) são uma ampliação da regressão linear clássica, possibilitando que a variável resposta siga distribuições distintas da normal, como a Binomial, Poisson e Gama (Cordeiro; Demétrio, 2008). Embora o MLG seja uma ferramenta valiosa para analisar dados complexos, oferece maior flexibilidade e robustez, também apresenta limitações, como maior esforço computacional, interpretação mais técnica, necessidade de escolher adequadamente a distribuição e a função de ligação, além da sensibilidade a *outliers* e má especificação do modelo, ainda assim, destaca-se por sua capacidade de lidar com variáveis discretas e contínuas tornando bastante úteis em diferentes contextos aplicados (Frees, 2010).

O MLG é representado através da equação (3), e descreve a média da variável resposta transformada por uma função de ligação sendo determinada por múltiplas variáveis explicativas (Pérez, 2022).

$$g(\mu_i) = n_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \tag{3}$$

Onde:

$\mu_i = E(Y_i)$: è a média da variável resposta Y_i;

$g(.)$: é a função de ligação que relaciona a média da resposta ao preditor linear;

n_i : é o preditor linear, combinação linear dos parâmetros e variáveis explicativas;

$\beta_0, \ldots, \beta_p$: são os parâmetros;

$X_{i1}, \ldots, X_{ip}$: variáveis independentes.

Neste contexto, um meio bastante utilizado para modelar variáveis positivas e assimétricas é o MLG com distribuição gama e função de ligação logarítmica, principalmente em contextos nos quais os dados representam valores monetários ou de custos (Ossifo *et al.*, 2024). A distribuição Gama é conhecida por se ajustar bem a variáveis contínuas estritamente positivas e com cauda longa, enquanto a função logarítmica assegura previsões positivas e proporciona uma relação multiplicativa entre a média da resposta e as variáveis independentes, permitindo interpretações consistentes em termos de efeitos proporcionais (Denuit; Hainaut; Trufin, 2019).

O MLG com distribuição Gama e função de ligação log é representada pela equação (4), e descreve a média da variável resposta positiva e assimétrica transformada pelo logaritmo sendo determinada por várias variáveis explicativas (Pérez, 2022).

$$g(\mu_i) = log(\mu_i) = n_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \quad (4)$$

Onde:

$\mu_i = E(Y_i)$: é a média da variável resposta Y_i;

$g(.)$: é a função de ligação que relaciona a média da resposta ao preditor linear;

$log(\mu_i)$: é a transformação logarítmica da média;

n_i : é o preditor linear, combinação linear dos parâmetros e variáveis explicativas;

$\beta_0, \ldots, \beta_p$: são os parâmetros;

$X_{i1}, \ldots, X_{ip}$: variáveis independentes.

A abordagem de MLG com distribuição gama e função de ligação log é frequentemente utilizada na previsão da severidade dos sinistros no setor securitário, tendo em vista que a indenização do sinistro tende a ser menor, mas podem assumir valores elevados em determinadas ocorrências, o modelo gama-log captura de forma realista essa distribuição assimétrica (Frees, 2010).

2.4 MODELAGEM DE APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina (AM), igualmente conhecido como *Machine Learning*, é um ramo da Inteligência Artificial (IA) que possibilita aos sistemas computacionais aprenderem automaticamente com dados, sem a necessidade de programação explícita para realizar tarefas específicas, mediante a detecção de padrões em grandes volumes de dados, esses sistemas são capazes de realizar previsões, classificações e apoiar a tomada de decisão com elevada autonomia (James *et al.*, 2023).

Faceli *et al.* (2025) divide os modelos de AM em três grandes categorias: o aprendizado supervisionado, o não supervisionado e o aprendizado por reforço. O Aprendizado Supervisionado utiliza algoritmos treinados com o conjunto de dados que envolve entrada e saídas esperadas, permitindo ao modelo compreender as relações entre as variáveis para, em seguida, realizar previsões. No Aprendizado Não Supervisionado, as informações não dispõem de rótulos, e o modelo procura identificar padrões implícitos ou segmentos naturais nos dados. Por fim, o Aprendizado por Reforço engloba a interação de um agente com o ambiente, aprendendo mediante recompensas ou penalidades (Hariom; Brad, 2024).

Com base nesta classificação, a modelagem de AM refere-se ao processo de construção, treinamento, validação e implantação dos modelos baseado em dados, este processo engloba a aplicação de distintos tipos de aprendizado, que melhor se encaixa na natureza do problema e informações disponíveis, permitindo ao modelo generalizar e produzir previsões precisas (Hariom; Brad, 2024).

No setor securitário, a modelagem de aprendizado de máquina vem se tornando de grande relevância para as seguradoras, pois permite uma avaliação mais exata do risco, adaptação da precificação de seguros e a constatação de fraudes (Vandervorst; Verbeke; Verdonck, 2022). As seguradoras convivem com grandes volumes de dados históricos e operacionais, tornando esta área ideal para a utilização de modelos preditivos (Vinuesa *et al.*, 2025). Os modelos de regressão e modelos lineares generalizados são frequentemente usados para estimar a probabilidade de ocorrência de sinistros (Steiner; Meng, 2019). Além disso, técnicas avançadas, como árvore de decisão, *Random Forest* e XGBoost, são comumente utilizadas na segmentação de clientes e na avaliação de risco (James *et al.*, 2023).

Consoante a isto, a modelagem de AM no setor agropecuário vem ganhando destaque pela capacidade de prever safras, otimizar o uso de insumos, diagnosticar doenças em plantações e animais e prever variações climáticas, permitindo maior eficiência na produção e diminuição de perdas (Liakos *et al.*, 2018). A título de exemplo, as redes neurais podem ser

utilizadas na visão computacional para identificar pragas nas lavouras através da análise de imagens (Kamilaris; Prenafeta-Boldú, 2018), e o *Random Forest* é amplamente utilizado na previsão de rendimento (Ajzan *et al.*, 2023).

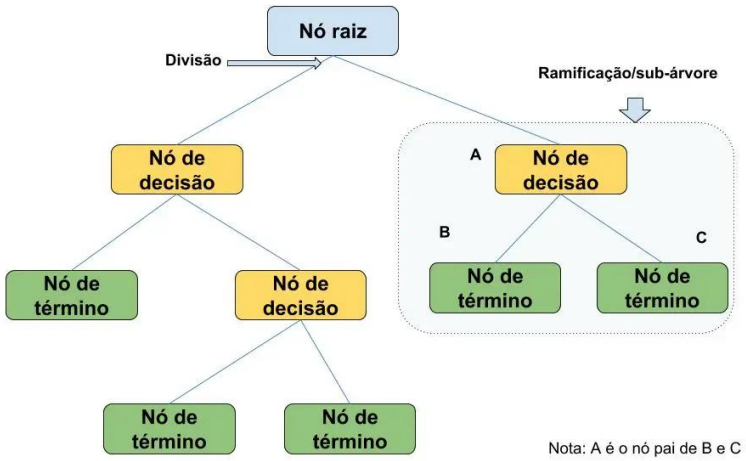
Segundo Vinuesa *et al* (2025), o AM viabiliza a identificação de padrões e efetua previsões com base em grandes volumes de dados, viabilizando impactos transformadores em setores securitários, em especial na agropecuária. Na agropecuária, os algoritmos de AM são aplicados para analisar dados climáticos, imagens de satélite e informações sobre o solo, ocasionando maior precisão no planejamento e diminuição de desperdícios (Liakos *et al.*, 2018). No entanto, Araújo *et al.* (2023) ressaltam que apesar das vantagens do AM, a sua utilização no setor agropecuário enfrenta desafios significativos, principalmente na falta de conjuntos de dados robustos e confiáveis, que englobam condições climáticas, do solo e de método de cultivo, essa limitação gera incertezas na modelagem e reduz a qualidade das previsões.

No setor securitário e agropecuário, os modelos de *Random Forest* e XGBoost vem sendo utilizado por sua capacidade de lidar com dados volumosos, capturar interações complexas entre variáveis e gerar previsões robustas, tornando-se uma ferramenta valiosa para avaliar o risco e detectar padrões relevantes nos dados (Ajzan *et al.*, 2023).

2.4.1 Modelos de AM *Random Forest* e XGBoost

Segundo Blockeel *et al.* (2023), os algoritmos de aprendizagem que utilizam a árvore de decisão estão entre os modelos de aprendizagem supervisionada mais eficazes e empregados. O autor conceitua a árvore de decisão como um método de análise e modelagem aplicada para auxiliar na tomada de decisão e resolução de problemas de forma estruturada, permitindo classificar informações, prever resultados e identificar padrões. A árvore de decisão funciona como um diagrama em formato de árvore, no qual cada nó interno (nó de decisão) representa uma condição ou teste, cada ramo indica um resultado possível desse teste e cada nó folha (nó de término) representa uma decisão ou previsão final, como demonstrado na Figura 2. Entre os modelos de árvores de decisão mais empregados, destaca-se o *Random Forest* e o XGBoost.

Figura 2 – Árvores de decisão.

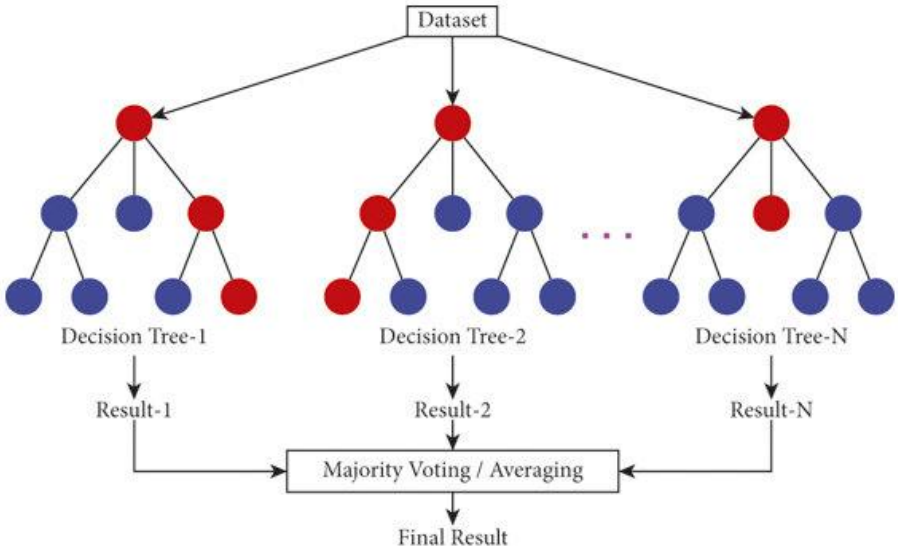


Fonte: Vooo (2016), adaptado de Analytics Vidhya.

O *Random Forest* (RF) é um algoritmo de AM supervisionado e não paramétrico do tipo *ensemble*, ou seja, eleva o desempenho preditivo ao combinar os resultados de vários algoritmos individuais (Zhou, 2012). O algoritmo constrói múltiplas árvores de decisão a partir de subconjuntos aleatórios dos dados e das variáveis explicativas, agregando-se os resultados dessas árvores, o modelo apresenta maior robustez, precisão e capacidade de generalização reduzindo significativamente o risco de *overfitting* (Breiman, 2001).

O modelo opera a partir da elaboração numerosa de árvores de decisão, cada árvore é treinada com um subconjunto aleatório dos dados com reposição, técnica conhecida como *bootstrap*, e um subconjunto aleatório de variáveis em cada divisão, introduzindo diversidade entre as árvores (Breiman, 2001). A predição final é adquirida por votação majoritária (*Majority Voting*), ou seja cada árvore vota por uma classe, e a mais votada define a predição final, nos casos de classificação, ou pela média das predições individuais (*Averaging*), a saída final é a média das predições feitas por todas as árvores, em problemas de regressão (Zhou, 2012).

Figura 3 – Ilustração do modelo Random Forest.



Fonte: Khan *et al.* (2021).

A Figura 3 simboliza o modelo *Random Forest*, no qual o *dataset* representa o conjunto de dados originais utilizados para o treinamento das várias árvores de decisão, neste modelo cada árvore recebe uma amostra diferente do conjunto original e seleciona aleatoriamente os subconjuntos das variáveis para os nós, além disso, cada árvore de decisão é treinada de forma independente (*decision tree* 1, 2 …n) e gera um resultado individual (*result* 1, 2 … n). Se o problema for de classificação utiliza-se a votação majoritária e se for de regressão utiliza-se as médias das previsões de todas as árvores (Khan *et al.*, 2021).

No setor securitário, o RF pode ser empregado para prever a severidade dos sinistros, no qual cada árvore é construída a partir de amostras aleatórias dos dados históricos de sinistros considerando subconjuntos aleatórios das variáveis explicativas, como idade do segurado, tipo de cobertura, região, histórico de sinistros, entre outros (Staudt; Wagner, 2021). A previsão final da severidade do sinistro é obtida através da média das previsões de todas as árvores por se tratar de valores contínuos, permitindo capturar relações complexas e não lineares entre as variáveis (Khan *et al.*, 2021).

A média das previsões é representada pela equação (5), e descreve a estimativa esperada da variável resposta dado o conjunto de variáveis explicativas (Khan *et al.*, 2021).

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{5}$$

Onde:

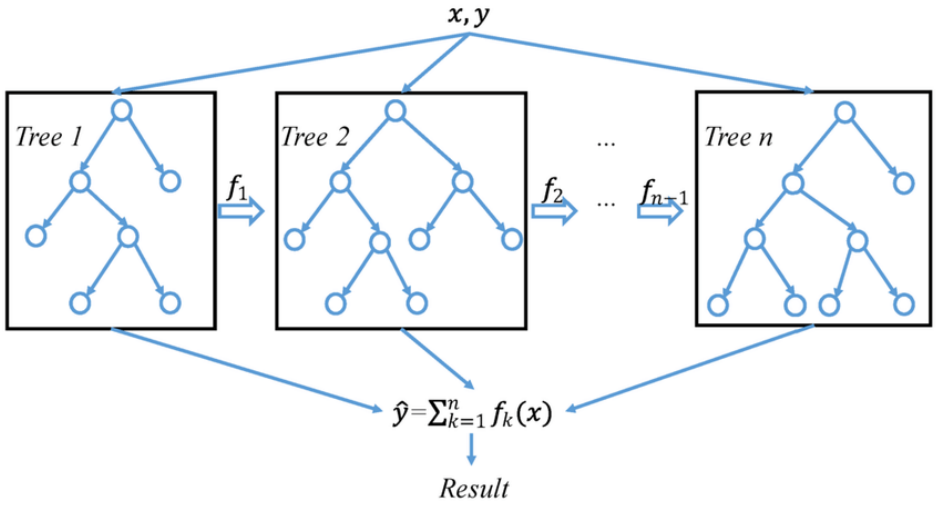
$\hat{y}$: é a previsão final do modelo para a observação x;
B: é o número total de árvores na floresta;
$T_b(x)$: é a previsão da b-ésima árvore para a observação x;
$\sum_{b=1}^{B} T_b(x)$: é a soma das previsões individuais de cada árvore T_b para a observação x.

A interpretação da equação (5) evidencia que a previsão final do *Random Forest* resulta da média das estimativas fornecidas por cada árvore de decisão. Esse processo de agregação reduz a variância do modelo e aumenta sua robustez, uma vez que erros individuais das árvores tendem a se compensar. Dessa forma, o Random Forest combina múltiplas previsões para produzir resultados mais estáveis e precisos (Staudt; Wagner, 2021; Khan *et al.*, 2021).

O modelo de *Random Forest* é utilizado por sua capacidade de lidar com grandes volumes de dados, reduzir a variância por meio da agregação de múltiplas árvores e apresentar bom desempenho em conjuntos com muitas variáveis, além de permitir a avaliação da relevância das variáveis envolvidas na predição. Contudo, também apresenta limitações, como menor desempenho em dados esparsos ou altamente dimensionais, baixa interpretabilidade, maior tempo de predição e tendência ao overfitting (Breiman, 2001). Além disso, outras abordagens, como o XGBoost, vêm ganhando destaque por sua eficiência e precisão (Ganie; Wani; Wani, 2026).

O XGBoost (*Extreme Gradient Boosting*), é um algoritmo de AM supervisionado e não paramétrico (Georganos *et al.*, 2018). Este algoritmo é estruturado em árvore de decisão, elaborado com foco em alto desempenho e eficiência computacional (Ganie; Wani; Wani, 2026). O modelo diferencia-se do *Random Forest* que utiliza múltiplas árvores independentes, pois constrói árvores dependentes, somando gradualmente os aprendizados de cada uma utilizando o método *boosting*, no qual árvores simples são implementadas de forma sequencial, cada um corrigindo o erro do anterior utilizando técnicas de regularização, paralelização e poda, tornando o modelo eficiente e preciso até em conjuntos de dados volumosos, ruidosos ou com alta dimensionalidade, contudo, apresenta limitações como complexidade e sensibilidade aos hiperparâmetros, menos interpretável em relação ao RF e maior risco de *overfitting* (Georganos *et al.*, 2018).

Figura 4 – Ilustração do modelo XGBoost.



Fonte: Wang *et al.* (2019).

A Figura 4 representa o modelo XGBoost, o x simboliza as variáveis de entrada e o y o valor esperado, cada caixa significa uma árvore de decisão, com isso a primeira árvore denominada de *tree* 1 é treinada com os dados de entrada x para prever y resultando na função f_1 , a segunda árvore (*tree* 2) é treinada para corrigir os erros da primeira árvore, ou seja, ela tenta modelar o que a primeira árvore não conseguiu prever corretamente, sendo representado pela função f_2, o processo continua até a n-ésima árvore (*tree* n), cada uma tentando melhorar a previsão combinando o que as anteriores não capturaram. As saídas de todas as árvores são somadas para formar a previsão final (Wang *et al.*, 2019).

Consoante a isto, o XGBoost, assim como o RF, também é utilizado para prever a severidade do sinistro, em que cada árvores é treinada para minimizar a diferença entre os valores reais dos sinistros e as previsões anteriores, permitindo capturar relações não lineares e interações complexas entre as variáveis explicativas (Fauzan; Murfi, 2018).

A soma das previsões de várias árvores treinadas sequencialmente é representada pela equação (6), e descreve o processo de construção aditiva, em que cada árvore complementa as anteriores (Fauzan; Murfi, 2018).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{6}$$

Onde:

$\hat{y}_i$: é a previsão final do modelo para a observação i;

K: é o número total de árvores de decisão que serão somadas;
$f_k(x_i)$: é a previsão da k-ésima árvore para a observação x_i;
$\sum_{k=1}^{K} f_k(x_i)$: é a soma sequencial das previsões de todas as árvores

Os modelos *Random Forest* e XGBoost são conhecidos e utilizados por vários setores por se mostrarem modelos voláteis, robustos e eficientes, no entanto, esses modelos são conhecidos por serem modelos que não são interpretáveis, ou seja, modelos difíceis de compreender como chegaram a determinado resultado utilizando determinadas variáveis, para solucionar este problema utiliza-se técnicas de interpretabilidade (Masís, 2023).

2.4.2 Técnicas de interpretabilidade e explicabilidade nos modelos de AM

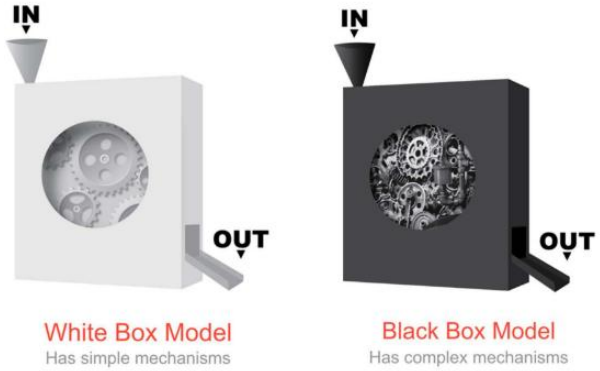
Atualmente, a crescente aplicação de modelos de AM em áreas como saúde, finanças e segurança tem aumentado a demanda por sistemas que não se restringem a gerir previsões precisas, mas também sejam compreensíveis e confiáveis (Rifaldi, 2025; Vinuesa *et al.*, 2025). Nesse contexto, emergem os conceitos de interpretabilidade e explicabilidade, no qual se referem à competência dos modelos ou sistemas explicar, concisamente, como se chegou a determinada decisão ou resultado (Masís, 2023).

Serg Masís (2023) conceitua e diferencia a interpretabilidade e a explicabilidade utilizadas em modelos de AM. Para o autor, a interpretabilidade está ligada à capacidade do ser humano compreender a lógica por trás da decisão do modelo, ou seja, porque uma determinada entrada gera uma saída específica ou quais variáveis mais influenciam a tomada de decisão dos modelos. Já a explicabilidade inclui tudo que a interpretabilidade abrange, no entanto, a explicabilidade exige a transparência sobre como o modelo funciona e como ele é treinado, ou seja, como ele foi construído, aprendeu com os dados e o seu funcionamento interno.

Os modelos de AM podem ser divididos em modelos intrinsecamente interpretáveis, “caixa branca” e modelos intrinsecamente não interpretáveis, “caixa preta” (Masís, 2023). Os modelos intrinsecamente interpretáveis ou “caixa branca”, são aqueles cuja estrutura, funcionalidade e parâmetro são facilmente compreendidos por humanos, sem a necessidade de utilizar técnicas externas para explicar como esses modelos tomam decisões. Os modelos de regressão linear, logística, árvore de decisão são considerados modelos intrinsecamente interpretáveis (Molnar, 2020). Já os modelos intrinsecamente não interpretáveis ou “caixa preta” são modelos não entendíveis, cuja compreensão é difícil de como transformam os dados

de entrada em resultado, sem compreender claramente o que acontece por dentro do modelo que fazem eles tomarem determinada decisão, os modelos de rede neural profunda, *Random Forest* e XGBoost são considerados modelos intrinsecamente não interpretáveis (Masís, 2023).

Figura 5 – Ilustração dos modelos caixa branca e caixa preta.



Fonte: Serg Masís (2023).

A Figura 5 ilustra os modelos que são intrinsecamente interpretáveis, "caixa branca" e não interpretáveis, "caixa preta", evidenciando os mecanismos simples dos modelos caixa branca quanto a sua interpretabilidade e os mecanismos complexos dos modelos caixa preta dificultando a interpretabilidade.

Para os modelos do tipo caixa-preta, aplicam-se técnicas de interpretabilidade e explicabilidade conhecidas como técnicas Pós-Hoc. Essas técnicas são utilizadas após o treinamento e a operação do modelo de AM, permitindo explicar e interpretar suas decisões, ou seja, não são incorporadas durante a construção ou o treinamento, mas introduzidas posteriormente com o objetivo de aumentar a transparência dos resultados. Entre as principais técnicas Pós-Hoc destacam-se: LIME, SHAP, Feature Importance e Partial Dependence Plots (Masís, 2023; Molnar, 2020).

Entre as técnicas de Pós-Hoc a mais destacada é a técnica SHAP (*SHapley Additive exPlanations*),representada pela equação (7), esta técnica funciona para modelos complexos, é consistente e baseado em teoria matemática sólida, permite explicar previsões individuais e o comportamento global do modelo e sua limitação está ligada ao custo computacional, (Molnar, 2020).

$$f(x) = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j \qquad (7)$$

Onde:

$f(x)$: é a previsão do modelo para a observação x;

ϕ_0: é o valor de referência (baseline);

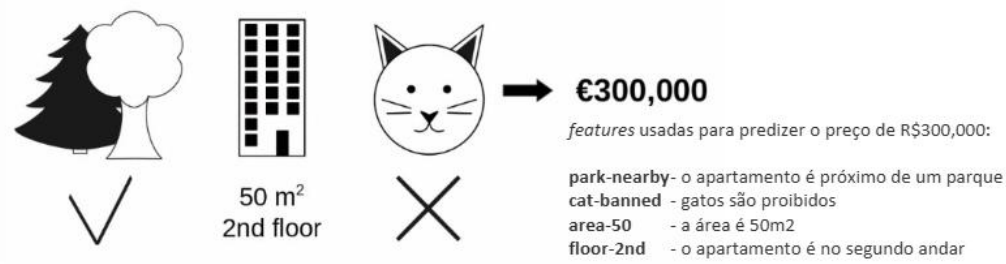
$\emptyset_j$: é o valor de Shapley associado à variável j;

M: é o número total de variáveis explicativas usadas pelo modelo;

$\sum_{j=1}^{M} \quad \emptyset_j$: é a soma das contribuições individuais de todas as variáveis explicativas do modelo.

Galvão (2024) conceitua o SHAP (*SHapley Additive exPlanations*) como uma técnica de interpretação de modelos de AM embasada nos valores de Shapley, um conceito originado na teoria dos jogos cooperativos, que o objetivo é explicar o resultado do modelo identificando a contribuição de cada variável para a previsão realizada. A lógica consiste em que cada variável é como um "jogador" em um jogo, e o objetivo é calcular, de forma justa, qual a participação de cada um no "resultado final" que a previsão. O autor destaca também que o SHAP avalia todas as combinações possíveis de variáveis e mede quanto a agregação de uma variável altera a previsão e o valor médio dessas alterações, considerando todos os cenários possíveis, representa a importância dessa variável para o resultado.

Figura 6 – Ilustração sobre as variáveis que influenciam a decisão da técnica SHAP.



Fonte: Galvão (2024).

A Figura 6 retrata a utilização do SHAP para saber qual variável teve maior impacto na previsão do valor do apartamento em 300.000 mil euros. A técnica empregada possibilitou a identificação que a variável *park-nearby* teve um impacto positivo de 15.000 mil euros e a variável *cat-banned* teve um impacto negativo em 30.000 mil euros, enquanto as variáveis área-50 teve um impacto positivo de 5.000 mil euros e *floor-2nd* não teve relevância para a previsão. Portanto, a *cat-banned* foi o fator de maior influência no preço estimado.

2.5 ESTUDOS ANTERIORES

A modelagem de risco no setor securitário tem evoluído com a incorporação do aprendizado de máquina e estatística avançada, especialmente em contextos complexos como o seguro agropecuário. A previsão de sinistros neste setor exige modelos capazes de lidar com dados complexos, variáveis interdependentes e padrões não lineares. Diversos estudos têm explorado abordagens estatísticas tradicionais, como a Regressão Linear Múltipla (RLM) e os Modelos Lineares Generalizados (MLG), além de técnicas modernas de aprendizado de máquina, como *Random Forest* e XGBoost, esses modelos têm sido aplicados tanto na previsão da frequência quanto na severidade dos sinistros, com resultados variados conforme o tipo de dado e o contexto operacional.

O autor Mahajan *et al.* (2025) investigou os aspectos da regulamentação europeia sobre a inteligência artificial na subscrição de seguros de vida e saúde, comparando modelos estatísticos tradicionais, como a Regressão Linear Múltipla (RLM), com algoritmos de aprendizado de máquina, como o XGBoost. Utilizaram uma base de dados com mais de 12 milhões de registros, observaram que os modelos de AM apresentaram maior precisão preditiva na estimativa de riscos de mortalidade, morbidade e cancelamento de apólices, enquanto a RLM, foi considerada mais transparente e estável, especialmente em ambientes regulatórios que exigem justificativas claras para decisões automatizadas.

Denuit, Hainaut e Trufin (2019) investigaram a aplicação de MLG em seguros pessoais e comerciais, com o objetivo de aprimorar a modelagem de frequência e severidade de sinistros por meio de distribuições flexíveis como Poisson e Gama. Os autores demonstraram que os MLGs são particularmente eficazes em bases de dados bem estruturadas, oferecendo robustez estatística e capacidade de adaptação a diferentes tipos de variáveis atuariais. Além disso,

Clemente *et al.* (2023) realizaram uma comparação entre MLGs e algoritmos de aprendizado de máquina, como *Gradient Boosting*, na previsão de sinistros automotivos. O estudo revelou que os MLGs apresentaram melhor desempenho na modelagem da severidade dos sinistros, enquanto os modelos de aprendizado de máquina foram superiores na previsão da frequência.

Consoante a isto, Brati *et al.* (2025) exploraram o uso de algoritmos de aprendizado de máquina, como *Random Forest* e XGBoost, na classificação de sinistros de alto custo em seguros automotivos. O estudo evidenciou a capacidade desses modelos em lidar com grandes volumes de dados e identificar padrões complexos, com destaque para o desempenho do *Random Forest*, que obteve acurácia de 88,67% e AUC de 0,9437, superando os demais

modelos avaliados. As variáveis com maior influência na classificação foram o tipo de sinistro, a marca do veículo e a idade do condutor, indicando a relevância de fatores comportamentais e técnicos na precificação de riscos.

No contexto da previsão da severidade, Su e Bai (2020) propuseram um modelo preditivo para a severidade de sinistros que incorpora dependência entre frequência e severidade, utilizando o algoritmo XGBoost aliado a uma abordagem de perfil de verossimilhança. Essa estrutura permitiu capturar interações não lineares e relações complexas entre variáveis, superando o desempenho dos MLGs tradicionais, que assumem independência entre frequência e severidade. O estudo evidenciou o potencial do XGBoost em contextos atuariais avançados, especialmente em situações aonde a estrutura dos dados exige maior flexibilidade e capacidade de modelagem de padrões ocultos.

Ademais, Noorunnahar *et al.* (2023) exploraram o uso do algoritmo XGBoost na previsão da produção agrícola de arroz em Bangladesh, demonstrando sua eficácia em cenários agropecuários. O modelo obteve um erro percentual absoluto médio (MAPE) de 5,38%, superando o desempenho do modelo ARIMA, que registrou MAPE de 7,23%. Esses resultados destacam o potencial do XGBoost para aplicações em seguro rural, especialmente na estimativa de variáveis que influenciam a precificação e o gerenciamento de risco climático.

Finalmente, Zhang (2024) desenvolveu métricas voltadas à avaliação da interpretabilidade de modelos de aprendizado de máquina, com destaque para o indicador *Signal Success Share* (Índice de Sucesso do Sinal), que quantifica a capacidade de um modelo explicar suas decisões de forma consistente e compreensível. O estudo teve como objetivo tornar algoritmos como *Random Forest* e XGBoost mais transparentes e adequados para uso em setores regulados, como o de seguros, onde decisões automatizadas precisam ser justificáveis. Os resultados mostraram que, quando combinados com ferramentas de interpretabilidade como SHAP, esses modelos podem atingir níveis satisfatórios de interpretação sem comprometer a acurácia preditiva.

Assim, o presente estudo visa ampliar os achados de autores como Mahajan *et al.* (2025), Denuit, Hainaut e Trufin (2019), Clemente *et al.* (2023) e Brati *et al.* (2025), integrando ainda as contribuições de Su e Bai (2020), Noorunnahar *et al.* (2023) e Zhang (2024), ao investigar, no âmbito do mercado securitário, a aplicação de modelos estatísticos e de aprendizado de máquina na previsão de severidade de sinistros, com especial atenção ao seguro agropecuário. Ao adotar uma abordagem empírica, o estudo pretende contribuir para a literatura ao oferecer visões mais práticas sobre o impacto das diferentes metodologias — como Regressão Linear Múltipla, Modelos Lineares Generalizados, *Random Forest* e XGBoost — na

modelagem de risco, na precificação atuarial e na conformidade regulatória, considerando aspectos como acurácia preditiva, interpretabilidade e adaptabilidade a dados complexos e não lineares.

3 METODOLOGIA

Nesta seção, será apresentado a trajetória metodológica estruturada com a finalidade de cumprir os objetivos estabelecidos neste trabalho. Será aplicado uma abordagem quantitativa e comparativa, direcionado na avaliação da eficácia de modelos estatísticos tradicionais e modelos interpretáveis de aprendizado de máquina na modelagem de risco do seguro agropecuário. O trabalho será realizado em etapas sistemáticas que englobam a coleta e tratamento de dados, aplicação de modelos e análise de desempenho.

As análises e comparações serão realizadas através do *software* Python (versão 3.13.7) , que disponibilizou as ferramentas necessárias para a execução dos objetivos propostos.

3.1 TIPO DE PESQUISA

Em relação à abordagem, o trabalho se enquadra na abordagem quantitativa, empregando modelos estatísticos e de aprendizado de máquina em conjunto com técnicas de interpretação e explicação para mensurar, modelar e comparar o desempenho preditivo dos modelos analisados.

Em relação à natureza, consiste em uma pesquisa aplicada, visto que busca produzir conhecimento voltado à solução de problemas práticos relacionados à modelagem de risco no seguro agropecuário.

Em relação aos objetivos, trata-se de uma pesquisa descritiva e comparativa, visando descrever o comportamento dos modelos estatísticos e de AM interpretáveis, além de comparar o desempenho preditivo e nível de interpretabilidade.

Em relação aos procedimentos técnicos, utiliza-se estudo de caso com modelagem computacional analisando dados reais de sinistros e aplicando modelos estatístico e de AM para a construção e avaliação dos modelos preditivos.

A delimitação temporal trata da análise de dados históricos compreendidos entre os anos de 2016 e 2024. Essa delimitação é importante, pois garante a consistência das análises e a validade dos resultados.

3.2 DADOS DA PESQUISA

Os dados utilizados para a realização dos cálculos estatísticos e de aprendizado de máquina serão extraídos do banco de dados, Sistema de Subvenção Econômica ao Prêmio do Seguro Rural (SISSER)[3] que é utilizado na operacionalização do Programa de Subvenção ao Prêmio do Seguro Rural (PSR), através de troca de informações entre o MAPA e as seguradoras habilitadas no programa, constando as informações referentes aos produtores que receberam a subvenção e os dados das apólices recepcionadas. Essa escolha foi fundamentada na alta qualidade, consistência e confiabilidade dos dados disponíveis, fornecendo elementos indispensáveis para a previsão da severidade dos sinistros por região. Cabe ressaltar que não serão utilizadas todas as variáveis disponíveis no banco de dados, mas apenas aquelas consideradas relevantes para alcançar o objetivo proposto, de modo a garantir maior precisão e eficiência na análise, na Tabela 1 foram consideradas as variáveis utilizadas.

O SISSER disponibiliza informações pertinentes, organizadas conforme apresentado na Tabela 1.

Tabela 1 – Informações sobre as variáveis relevantes do banco de dados SISSER.

Variáveis	Descrição
NM_MUNICIPIO_PROPRIEDADE	Nome do município onde está localizada a propriedade.
SG_UF_PROPRIEDADE	Sigla da Unidade da Federação onde está localizada a propriedade.
NM_CLASSIF_PRODUTO	Classificação do tipo de seguro
NM_CULTURA_GLOBAL	Cultura ou atividade segurada.
NR_AREA_TOTAL	Área total segurada.
NR_ANIMAL	Número de animais segurados.
NR_PRODUTIVIDADE_SEGURADA	Produtividade segurada.
NivelDeCobertura	Nível de cobertura do seguro.
VALOR_INDENIZAÇÃO	Valor pago em indenização, em caso de sinistro.
EVENTO_PREPONDERANTE	Evento preponderante causador do sinistro.
VL_LIMITE_GARANTIA	Limite de garantia da apólice (exposição financeira do contrato).
NM_CLASSIF_PRODUTO	Classificação do produto segurado (agrícola ou pecuário).
ANO_APOLICE	Ano de contratação da apólice (utilizado como variável indicadora/dummy no modelo).

Fonte: Elaboração própria, com base nos dados disponibilizados pela CGSEG (2025).

[3] Unidade Responsável pela Base de Dados: Coordenação-Geral de Seguro Rural (CGSEG). Disponível em: https://dados.agricultura.gov.br/dataset/sisser3

As variáveis Limite de Garantia (VL_LIMITE_GARANTIA), Classificação do Produto (NM_CLASSIF_PRODUTO) e Ano da Apólice (ANO_APOLICE) integram o conjunto de variáveis utilizadas nos modelos finais. O limite de garantia é, inclusive, o preditor de maior efeito em todos os modelos estimados (Seção 4.4). A variável Região (REGIAO), derivada de SG_UF_PROPRIEDADE, também é utilizada nos modelos, mas não constitui variável primária do banco SISSER, razão pela qual não recebeu linha própria na tabela. VL_PREMIO_LIQUIDO e VL_SUBVENCAO_FEDERAL constam na tabela por completude do banco de dados, mas foram excluídas da modelagem final por multicolinearidade severa (Seção 4.1), conforme assinalado nas respectivas linhas.

Os valores monetários foram deflacionados para valores reais de 2024, representada pela equação (8) (Neto, 2022), utilizando o índice IPCA acumulado, com o objetivo de eliminar os efeitos da inflação sobre as variáveis monetárias, garantindo comparabilidade temporal entre os dados e evitando viés nas estimativas dos modelos preditivos. As informações do IPCA acumulado foram retiradas do site do SIDRA tabela 1737.

$$Valor\ corrigido\ =\ Valor\ nominal\ *\ \frac{índice\ base}{índice\ do\ ano} \tag{8}$$

3.3 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento de dados é crucial para assegurar a qualidade e a confiabilidade dos modelos preditivos. Este processo tem o objetivo de preparar a base de dados para a aplicação dos modelos estatísticos e de AM, garantindo a consistência, completude e adequação às técnicas de modelagem (Silva *et al.*, 2016).

Inicialmente, foi realizada a limpeza dos dados, que incluiu a identificação e o tratamento de valores ausentes, duplicatas e inconsistências. Para variáveis numéricas, valores ausentes foram imputados por meio da mediana, devido à sua robustez frente a *outliers*, enquanto para variáveis categóricas foi utilizado a moda e os registros duplicados ou inconsistentes foram removido e tratado, para evitar viés na modelagem (Silva *et al.*, 2016).

A mediana é representada pela equação 9, e a moda é representada pela equação 10 (Morettin; Bussab, 2010).

$$Med = \begin{cases} \frac{n+1}{2}, se\ n\ for\ impar \\ \frac{\frac{n}{2} + \left(\frac{n}{2} + 1\right)}{2}, se\ n\ for\ par \end{cases} \quad (9)$$

Onde:

x: é o valor da observação do conjunto de dados;
n: é o número total de observações no conjunto de dados.

$$Mod = arg_{xi} f(x_i) \quad (10)$$

Onde:

x_i: é cada valor observado no conjunto de dados;
$f(x_i)$: é a frequência absoluta (ou número de ocorrências) do valor x_i;
$arg_{xi} f(x_i)$: indica o valor de x_i que maximiza a frequência $f(x_i)$.

Subsequente, foi realizado a seleção e transformação de variáveis, nos quais, variáveis com baixa variância ou alta correlação foram identificadas por meio de análise de matriz de correlação e pelo fator de inflação da variância (FIV) e da avaliação de variância dos atributos (filtro de baixa variância), sendo excluídas para reduzir a redundância e multicolinearidade (Pyle, 1999). As variáveis categóricas foram codificadas utilizando a técnica *target encoding*, transforma cada categoria em um valor numérico baseado na média da variável alvo, garantindo a compatibilidade com os algoritmos, para variáveis numéricas com distribuição assimétrica, será aplicado a transformação logarítmica, visando aproximar sua distribuição à normalidade (Klugman *et al.*, 2019).

A matriz de correlação é representada pela equação 11 (Gujarati; Porter, 2011):

$$r_{jk} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} \left(x_{ij} - \underline{x}_j\right)\left(x_{ik} - \underline{x}_k\right)}{\sigma X_j \sigma X_k} \quad (11)$$

Onde:

x_{ij}, x_{ik}: variáveis preditoras;

$\underline{x}_j, \underline{x}_k$: médias;

$\sigma X_j \sigma X_k$: desvios-padrão amostral;

n : número de observações.

A matriz de correlação (equação 11) foi utilizada para identificar relações lineares entre as variáveis preditoras, a fim de detectar e tratar problemas de multicolinearidade. A análise dos coeficientes de correlação permitirá selecionar variáveis com maior poder explicativo para a severidade do sinistro, eliminando redundâncias e garantindo que o modelo final seja baseado em preditores mais robustos e independentes (Gujarati; Porter, 2011).

O FIV é representado pela equação 12 (Gujarati; Porter, 2011):

$$FIV_j = \frac{1}{1 - {R^2}_j} \tag{12}$$

Onde:

${R^2}_j$: coeficiente de determinação ao regredir x_j sobre todas as demais variáveis. Valores altos de FIV indicam multicolinearidade.

O FIV (equação 12) foi utilizado como métrica complementar à matriz de correlação para diagnosticar com precisão a multicolinearidade entre as variáveis preditoras. Ao identificar e remover variáveis com FIV elevado, será garantido maior robustez e confiabilidade aos modelos estatísticos tradicionais, assegurando que suas previsões não sejam comprometidas por redundâncias no conjunto de dados (Gujarati; Porter, 2011).

A avaliação de variância dos atributos é representada pela equação 13 (Morettin; Bussab, 2010).

$$Var(X_j) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \underline{x}_j\right)^2 \rightarrow \ se\ Var(X_j) \ < \ \epsilon, remover \ \ X_j \tag{13}$$

Onde:

X_j : variável preditora;

$\underline{x_j}$: média;

ϵ : limiar pré-definido;

n : número de observações.

A avaliação de variância dos atributos (equação 13) foi utilizada como um filtro para eliminar variáveis preditoras com baixa variabilidade, ou seja, que quase não se alteram entre as observações. Ao remover tais variáveis, o conjunto de dados é simplificado, melhora a eficiência do treinamento dos modelos e aumenta a relevância estatística das variáveis restantes, resultando em um modelo preditivo mais enxuto e generalizável.

A técnica *target encoding* é representada pela equação 14 (Pargent, 2022):

$$Encoding(x) = \frac{\sum_{i=1}^{n} 1\,(X_i = x) * Y_i + \alpha * \mu}{N_x + \alpha} \tag{14}$$

Onde:

X : variável categórica;

x : uma categoria específica de X;

Y_i : valor do alvo (target) na observação;

N_x: número de observações na categoria x;

μ: média global do alvo em todo o conjunto de dados;

α : parâmetro de suavização (hiperparâmetro).

A técnica de *Target Encoding* (equação 14) foi utilizada para transformar variáveis categóricas em valores numéricos baseados na média da variável alvo. Para reduzir o risco de *overfitting*, aplicou-se uma versão suavizada que combina a média da categoria com a média global do alvo, evitando valores extremos em grupos raros. Essa abordagem preserva a informação relevante das categorias, reduz a dimensionalidade e aumenta a robustez das previsões de severidade de sinistros (Pargent, 2022).

A transformação logarítmica é representada pela equação 15 (Morettin; Bussab, 2010).

$$X' = log(1 + X) \tag{15}$$

Onde:

X': valor transformado;
X: valor original da variável contínua;
$(1 + X)$: soma de 1 ao valor original, para evitar problemas quando X=0;
log : logaritmo natural (base e).

A transformação logarítmica (Equação 15) será aplicada a variáveis numéricas com distribuição assimétrica, que tipicamente apresenta uma concentração de valores baixos e uma cauda longa de valores elevados. Essa transformação tem como objetivo aproximar a distribuição dessas variáveis à normalidade, estabilizar a variância dos resíduos e reduzir a influência de *outliers*, melhorando assim o ajuste e a performance de modelos.

Adicionalmente, as variáveis numéricas contínuas foram padronizadas pelo método *StandardScaler* (biblioteca *Python*), que transforma os dados para média 0 e desvio padrão 1, favorecendo a convergência dos algoritmos e a comparabilidade dos coeficientes nos modelos estatísticos (James *et al.*, 2023). Embora os modelos baseados em árvores de decisão, como o *Random Forest* e o XGBoost, não sejam sensíveis à escala das variáveis, a padronização foi aplicada de forma uniforme a todos os modelos dentro do fluxo de validação, sem prejuízo para os algoritmos de árvore.

A *StandarScaler* é representada pela equação 16 (James *et al.*, 2023).

$$X'_i = \frac{X_i - \mu}{\sigma} \tag{16}$$

Onde:

X'_i: valor padronizado (média 0, desvio padrão 1);
X_i: valor original da variável;
μ : média da variável;
σ : desvio padrão da variável.

A técnica de *StandardScaler* Equação 16 foi empregada para padronizar variáveis numéricas, transformando-as em uma distribuição com média zero e desvio padrão unitário. Essa abordagem é particularmente benéfica para modelos estatísticos tradicionais, como a regressão linear múltipla e os MLGs, pois elimina viés de escala, facilita a comparação direta

entre coeficientes e otimiza a convergência de algoritmos baseados em gradiente. Ao centralizar e reduzir a variabilidade dos dados, melhora a estabilidade numérica e a interpretabilidade dos parâmetros estimados, assegurando que o modelo capture relações verdadeiras entre variáveis preditoras e a severidade do sinistro, sem distorções causadas por diferenças de magnitude.

No ambiente Python, a modelagem estatística e de aprendizado de AM foi conduzida através das bibliotecas *pandas* e *numpy* empregadas para manipulação eficiente e pré-processamento dos dados, garantindo que o conjunto esteja adequadamente estruturado para análise (Mckinney, 2025; NUMPY DEVELOPERS, 2023).

Para os modelos de Regressão Linear Múltipla (RLM) e Modelos Lineares Generalizados (MLG) com distribuição gama e função de ligação logarítmica, foi utilizada a biblioteca *statsmodels*, que oferece precisão estatística, estimadores robustos e diagnósticos detalhados (Seabold; Perktold, 2010).

Os algoritmos de *Random Forest* e XGBoost foram aplicados para explorar abordagens mais flexíveis e não lineares, permitindo maior capacidade preditiva. O *scikit-learn* forneceu ferramentas para treinar, ajustar e avaliar o *Random Forest* (SCIKIT-LEARN DEVELOPERS, 2025), enquanto o XGBoost disponibilizou uma implementação otimizada do algoritmo *Gradient Boosting*, garantindo eficiência computacional (XGBOOST DEVELOPERS, 2023).

3.4 MODELOS PREDITIVOS

Os modelos preditivos são ferramentas estatísticas e computacionais que permitem estimar valores futuros ou desconhecidos a partir de dados históricos e variáveis explicativas (Verbeke; Dejaeger; Martens, 2021). No contexto do presente estudo, tais modelos são empregados para prever a severidade dos sinistros, fornecendo subsídios relevantes para a precificação de produtos, análise de riscos e tomada de decisão no setor securitário.

No método estatístico, foram utilizados a Regressão Linear Múltipla (RLM), com a variável dependente transformada por logaritmo natural na forma log(1+y) (equação 2), e o Modelo Linear Generalizado (MLG) com distribuição gama e função de ligação logarítmica (equação 4), ambos implementados em Python por meio da biblioteca *statsmodels*.

Para o método de aprendizado de máquina, foram aplicados os modelos de *Random Forest* (equação 5) e XGBoost (equação 6). No *Random Forest*, implementado pelo *scikit-learn*, o processo de *grid search* explorou diferentes combinações de hiperparâmetros, incluindo o número de árvores (200, 300 e 400), a profundidade máxima (10, 15 e 18), o mínimo

de amostras necessárias para realizar uma divisão (10, 15 e 20), o mínimo de amostras por folha (5, 10 e 15) e o número máximo de atributos considerados em cada divisão (sqrt e log2). Já no XGBoost, versão otimizada do *Gradient Boosting*, a busca envolveu o número de árvores (200, 300, 400), a taxa de aprendizado (*learning rate*) com valores de 0,05, 0,1 e 0,2, a profundidade máxima (4, 6 e 8), além dos parâmetros de subamostragem (*subsample*) e de proporção de colunas (*colsample*), ambos avaliados nos valores 0,2, 0,5 e 0,8.Em ambos os casos, foi empregada a técnica SHAP para interpretabilidade, permitindo compreender a contribuição de cada variável nas previsões (Lundberg *et al.*, 2023).

A avaliação dos modelos foi conduzida por validação cruzada, com a estratégia KFold em partições embaralhadas e semente aleatória fixa, adotando-se dez partições para a RLM e cinco para o MLG Gama, o Random Forest e o XGBoost, em razão do custo computacional do reajuste completo do pré-processamento em cada partição. Em cada partição, o pré-processamento (transformação log(1+x), padronização e target encoding) foi ajustado apenas sobre os dados de treino e aplicado ao conjunto de teste, evitando o vazamento de informação.

Nesse contexto, a utilização combinada de técnicas estatísticas e de aprendizado de máquina permite não apenas comparar desempenhos, mas também fornecer diferentes perspectivas para a previsão da severidade de sinistros. Essa abordagem comparativa possibilita equilibrar precisão preditiva e capacidade de interpretação, fator essencial no setor securitário.

3.5 TÉCNICA DE INTERPRETABILIDADE

A interpretabilidade dos modelos “caixa-preta” – *Random Forest* e XGBoost – é um fator importante no contexto do seguro, pois além da preditividade, é necessário compreender como e por que determinadas variáveis influenciam o resultado da previsão (Molnar, 2020; Masís, 2023). Esse aspecto é especialmente relevante em ambientes regulados, em que decisões precisam ser justificadas de forma clara e transparente (Rifaldi, 2025; Vinuesa *et al.*, 2025).

Para este estudo, será utilizada a técnica SHAP (Lundberg *et al.*, 2023) como principal ferramenta de interpretabilidade, na qual, é representada pela equação (7), contribuindo no equilíbrio da acurácia preditiva e a interpretabilidade, garantindo que os resultados obtidos pelos modelos possam ser interpretados por especialistas do setor securitário, gestores e órgãos reguladores.

3.6 MÉTRICAS DE AVALIAÇÃO

A avaliação do desempenho dos modelos preditivos será realizada por meio de métricas estatísticas. Essas métricas permitem mensurar a precisão das previsões, comparando os valores estimados pelos modelos com os valores observados de severidade dos sinistros. Neste estudo, serão adotadas as seguintes métricas de erro: *Mean Absolute Error* (MAE) – Erro Absoluto Médio –, *Root Mean Squared Error* (RMSE) – Raiz do Erro Quadrático Médio – e o Erro Percentual Absoluto Médio Simétrico – *Symmetric Mean Absolute Percentage Error* (sMAPE).

Os modelos serão avaliados no conjunto de teste, ou seja, nos dados não utilizados durante o treinamento. Essa prática é essencial para verificar a capacidade de generalização e a robustez preditiva em dados não vistos, evitando estimativas excessivamente otimistas. Conforme Hyndman e Athanasopoulos (2018), a avaliação em dados independentes do treinamento é indispensável para medir o desempenho real de modelos preditivos e garantir validade estatística.

O MAE calcula a média das diferenças absolutas entre os valores observados e os valores previstos. Essa métrica expressa, em unidades da variável resposta, o erro médio cometido pelo modelo, sendo facilmente interpretada. Quanto menor o MAE, melhor o desempenho do modelo (James *et al.*, 2023). O MAE pode ser representado pela equação (17).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{17}$$

Onde:

y_i: representa o valor real;
$\hat{y}_i$: o valor previsto;
n: o número total de observações.

O RMSE é obtido pela raiz quadrada da média dos erros quadráticos. Por elevar ao quadrado as diferenças entre valores observados e previstos, penaliza mais fortemente erros elevados, sendo útil para avaliar a robustez do modelo em relação a *outliers* (James *et al.*, 2023). O RMSE pode ser representado pela equação (18).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (18)$$

Onde:

y_i: representa o valor real;
$\hat{y}_i$: o valor previsto;
n: o número total de observações.

O sMAPE é uma métrica utilizada para avaliar o desempenho de modelos preditivos em termos percentuais, apresentando uma formulação simétrica que equilibra superestimativas e subestimativas. Quanto menor o valor do sMAPE, mais próximas as previsões estão dos valores observados, indicando maior precisão do modelo (Maiseli, 2019). O sMAPE pode ser representado pela equação (19).

$$sMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \quad (19)$$

Onde:

y: representa o valor real;
$\hat{y}$: o valor previsto pelo modelo;
n : o número total de observações.

3.7 ANÁLISE COMPARATIVA

A etapa final da metodologia consiste na análise comparativa entre os modelos estatísticos tradicionais e os modelos de aprendizado de máquina interpretáveis aplicados à previsão da severidade dos sinistros. O objetivo dessa etapa é verificar qual abordagem apresenta maior eficácia em termos de precisão preditiva e interpretabilidade, considerando as particularidades do setor securitário agropecuário.

Para a comparação, os modelos regressão linear múltipla e modelos lineares generalizados serão confrontados com os algoritmos de *Random Forest* e XGBoost, de modo a

avaliar as diferenças de desempenho entre técnicas estatísticas clássicas e métodos mais modernos de aprendizado de máquina.

A avaliação será realizada por meio das métricas descritas no tópico anterior, aplicadas de forma uniforme a todos os modelos. Essa padronização garante equidade na análise, possibilitando identificar:

- O modelo com menor MAE, indicando previsões mais consistentes;
- O modelo com menor RMSE, indicando maior robustez frente a grandes desvios;
- O modelo com menor sMAPE, indicando que as previsões estão proporcionalmente mais próximas dos valores reais.

Além da comparação quantitativa, será conduzida uma análise qualitativa por meio da aplicação da técnica de interpretabilidade SHAP nos modelos de AM, permitindo identificar as variáveis de maior influência na previsão e verificar se os modelos de aprendizado de máquina fornecem explicações compatíveis com os fatores de risco reconhecidos pelo setor securitário. Além disso, para os modelos estatísticos RLM e MLG, a interpretabilidade será analisada por meio da avaliação direta dos coeficientes, seus sinais, magnitudes, significância estatística e estabilidade das estimativas, considerando a relação explícita entre preditores e a severidade do sinistro. A comparação entre os modelos de AM interpretável e modelos estatísticos será realizada com base na clareza, profundidade e consistência das explicações, permitindo identificar qual abordagem fornece maior transparência para a tomada de decisão no setor securitário.

Para os modelos estatísticos (RLM e MLG Gama), a significância dos coeficientes foi avaliada por meio do p-valor, adotando-se o nível de significância de 5% ($\alpha = 0{,}05$) como critério de decisão: coeficientes com p-valor inferior a esse limiar foram considerados estatisticamente diferentes de zero. Cabe destacar que a significância estatística indica apenas que o efeito estimado é distinto de zero com determinado grau de confiança, não devendo ser confundida com a magnitude ou a relevância prática (explicabilidade) da variável, que é avaliada separadamente pelo tamanho do coeficiente (efeito multiplicativo) e, nos modelos de aprendizado de máquina, pelos valores de SHAP.

Assim, a análise comparativa permitirá não apenas selecionar o modelo mais preciso, mas também avaliar o equilíbrio entre desempenho preditivo e interpretabilidade.

4 RESULTADOS

4.1 ANÁLISE EXPLORATÓRIA DOS DADOS

O conjunto original da base de dados do SISSER contém 1.048.566 registros e 38 variáveis, em que cada observação corresponde a uma apólice contratada, com informações sobre o produtor, o produto, a localização e os valores do seguro.

Nem todas as variáveis se mostraram pertinentes ao objetivo do estudo. A seleção considerou o significado e a relação de cada variável explicativa com a variável resposta, resultando em um subconjunto voltado à modelagem da severidade. A variável resposta adotada foi o valor da indenização deflacionado pelo IPCA, que expressa a severidade do sinistro em termos monetários reais. As variáveis explicativas reúnem características da produção e do contrato, como área total, número de animais, produtividade segurada, nível de cobertura, limite de garantia, cultura, evento preponderante, município, unidade da federação, região e ano de contratação.

O tratamento dos dados envolveu a correção de inconsistências, a conversão de variáveis numéricas registradas como texto e a substituição do separador decimal. Os valores ausentes concentraram-se em variáveis de natureza estrutural: o evento preponderante e o valor da indenização não estavam preenchidos em 71,25% dos registros, proporção que corresponde às apólices sem ocorrência de sinistro; o número de animais apresentou 25,56% de ausências, por não se aplicar às modalidades agrícolas; e o nível de cobertura, 6,39%. As demais variáveis explicativas estavam praticamente completas. Os campos ausentes foram imputados pela mediana, no caso das variáveis numéricas, e pela moda, no caso das categóricas. Nos registros sem ocorrência de sinistro, o valor da indenização foi fixado em zero e o evento preponderante recebeu a marcação correspondente à ausência de sinistro. Para a etapa de severidade, manteve-se apenas o subconjunto com indenização positiva, o que resultou em 190.423 observações, equivalentes a 18,2% das apólices da base completa.

A Tabela 2 apresenta a estatística descritiva das variáveis numéricas e da variável resposta, em escala original.

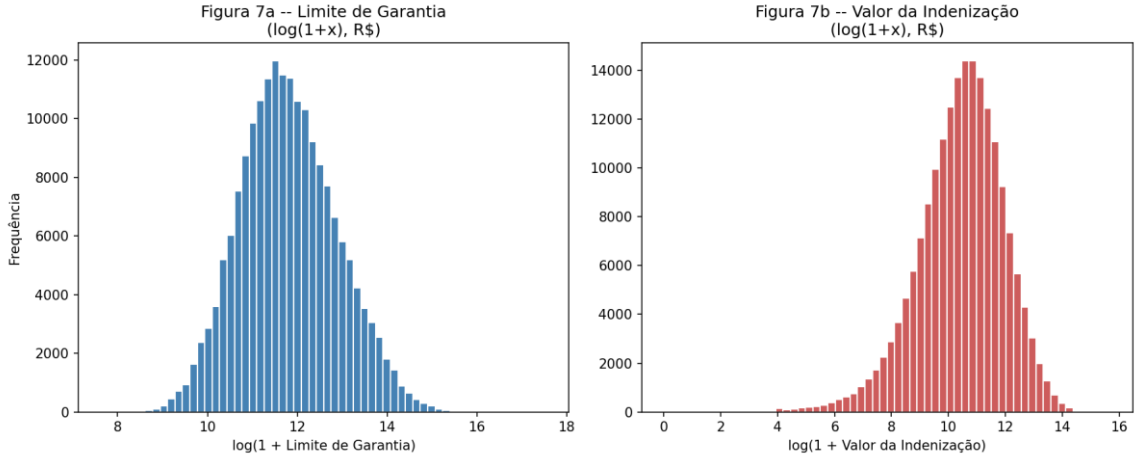
Tabela 2 – Estatística descritiva das variáveis (base de severidade, escala original).

Variável	Média	Mediana	DP	Mínimo	Máximo	Assimetria
Área total (ha)	60,02	29,00	702,63	0,00	291 Mil	382,18
Número de animais	0,27	0,00	15,48	0,00	4,1 Mil	149,49
Produtividade segurada	15,1 Mil	2,7 Mil	505,3 Mil	0,00	84 Milhões	79,26
Nível de cobertura	0,73	0,65	0,13	0,00	1,00	1,28
Limite de garantia (R$)	294,6 Mil	151,5 Mil	488,9 Mil	2,6 Mil	48,5 Milhões	15,09
Valor da indenização (R$)	101,1 Mil	45,8 Mil	173,7 Mil	0,01	8,6 Milhões	6,33

Fonte: Elaboração própria (2026).

Observa-se forte assimetria à direita na variável de severidade. A média do valor da indenização (R$ 101.119,77) supera a mediana (R$ 45.774,91) em cerca de 2,2 vezes, e o coeficiente de assimetria de 6,33 confirma a concentração de eventos extremos em uma parcela reduzida das observações. Esse padrão é característico do seguro agrícola, em que a maior parte das apólices apresenta perdas moderadas e poucos eventos climáticos severos produzem indenizações elevadas (Caffagni; Paixão; Rios, 2022). As variáveis explicativas contínuas exibem comportamento semelhante, com máximos muito distantes da mediana, o que justifica a transformação logarítmica adotada na modelagem e o uso de modelos capazes de lidar com não linearidade e caudas pesadas.

Figura 7 – Histogramas do limite de garantia e do valor da indenização (log(1+x)), apresentados separadamente.



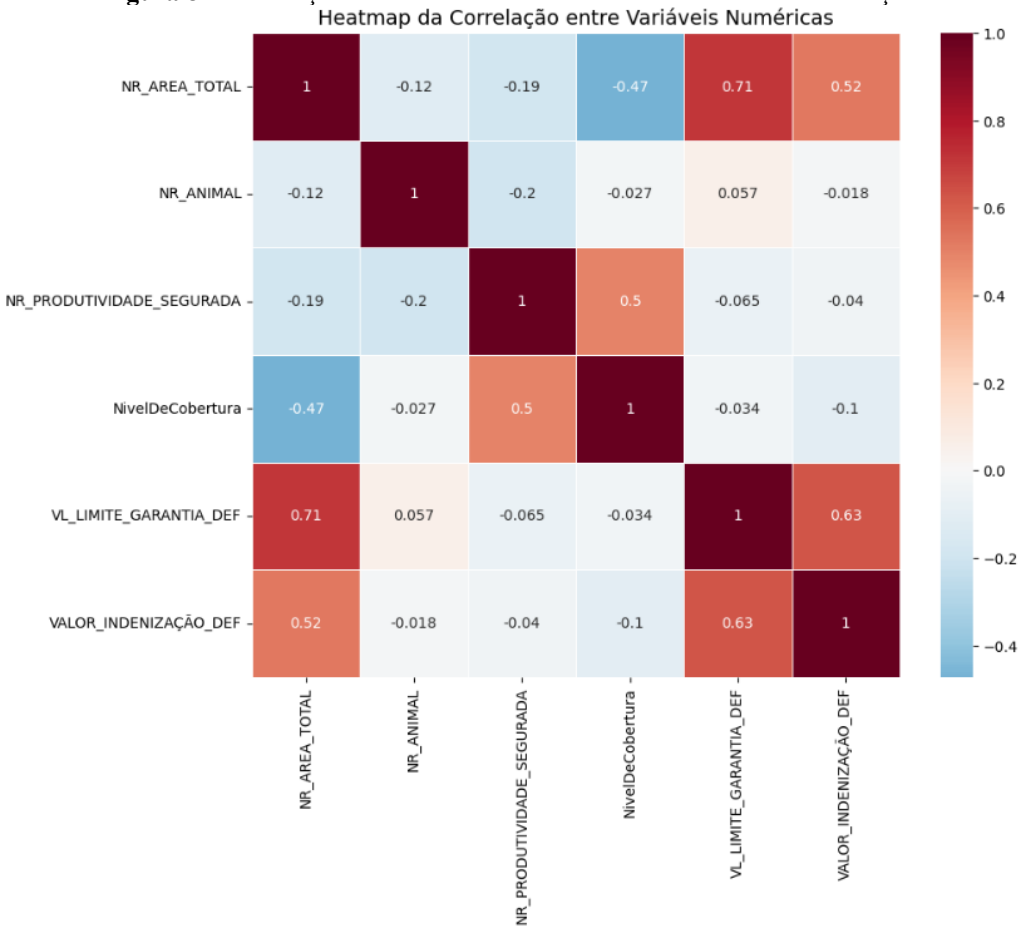
Fonte: Elaboração própria (2026).

A Figura 7 apresenta, em painéis separados, os histogramas do limite de garantia e do valor da indenização, ambos na escala log(1+x). Em escala monetária original, as duas variáveis concentram quase toda a massa de dados próxima à origem, em razão da forte assimetria à

direita e da presença de valores extremos, o que dificulta a leitura da distribuição. Aplicada a transformação logarítmica, cada variável revela um formato aproximadamente simétrico, comportamento coerente com o padrão de perdas do seguro agrícola, em que sinistros severos são pouco frequentes, porém de alto custo (Caffagni; Paixão; Rios, 2022).

Cabe registrar que não há valores nulos ou iguais a zero no limite de garantia, cujo mínimo observado na base bruta é de R$ 2.034,74; a concentração aparente próxima da origem decorre apenas do efeito de escala descrito acima, e não de falha de preenchimento do banco de dados. Para melhorar essa análise, a Figura 8 apresenta as correlações entre as variáveis numéricas e a variável resposta.

Figura 8 – Correlação entre as variáveis numéricas e o valor da indenização

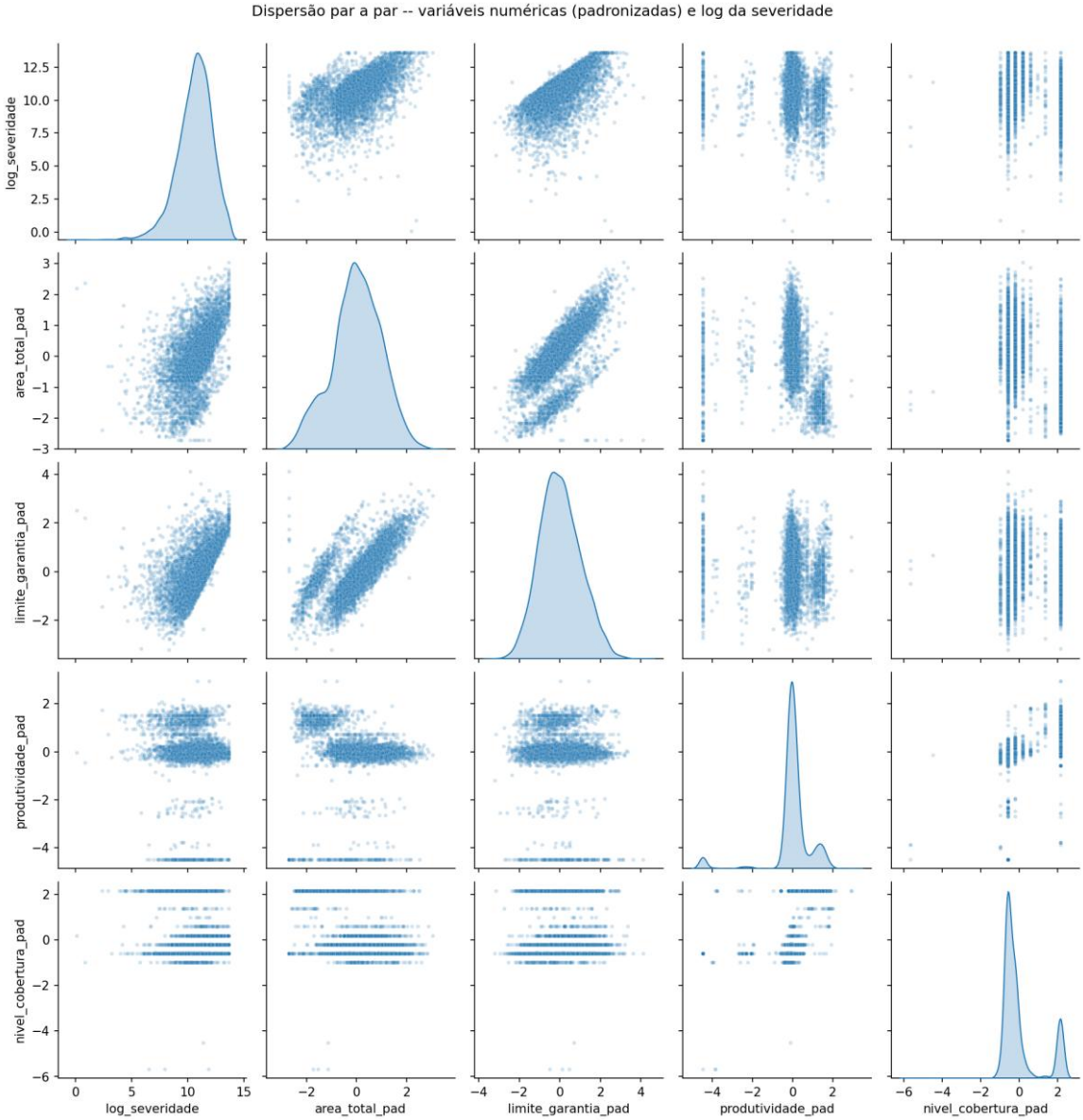


Fonte: Elaboração própria (2026).

A figura 8 nos mostra um limite de garantia (0,63) e a área total (0,52) sendo as maiores correlações com o valor da indenização, enquanto as demais variáveis contínuas mostram associação fraca. O resultado antecipa os principais determinantes da severidade, posteriormente confirmados pelos modelos preditivos.

Complementarmente ao heatmap de correlação (Figura 8), a Figura 9 apresenta a dispersão par a par (pairplot) das principais variáveis numéricas padronizadas com o log da severidade, o que permite identificar não linearidades e a discretização do nível de cobertura, variável percentual com poucos patamares distintos, padrões não visíveis no heatmap. Investigou-se também a correlação dos erros, potencialmente mais esclarecedora que a correlação entre as variáveis brutas para diagnosticar a heterocedasticidade.

Figura 9 - Dispersão par a par (pairplot) das variáveis numéricas padronizadas e do log da severidade.



Fonte: Elaboração própria (2026).

Quanto à correlação dos erros, calculou-se a correlação entre o quadrado dos resíduos da RLM, utilizado como aproximação da variância local do erro, e cada preditor, métrica mais

informativa que a correlação das variáveis brutas para investigar a heterocedasticidade. Os maiores coeficientes foram observados para o nível de cobertura (0,109), o evento preponderante (−0,106) e o indicador do ano de 2021 (−0,094), indicando que a dispersão dos erros varia conforme o nível de cobertura contratado, o tipo de evento causador do sinistro e o ano de estiagem severa, aspecto retomado na análise de heterocedasticidade da Seção 4.3.

Antes da modelagem, as variáveis contínuas foram transformadas por log(1+x), as variáveis numéricas foram padronizadas (média zero e desvio-padrão unitário) e a variável resposta foi truncada no percentil 99 para conter a influência de valores atípicos. As variáveis categóricas foram convertidas por *target encoding*, técnica em que cada categoria é substituída pela média da severidade observada naquela categoria (a média condicional do alvo).

Essa escolha se justifica pela alta cardinalidade de variáveis como o município, que inviabilizaria a codificação binária por elevar excessivamente a dimensionalidade; o *target encoding* comprime a informação em um único preditor numérico e, com o uso de suavização, aproxima as categorias raras da média global, reduzindo o risco de sobreajuste (Pargent *et al.*, 2022). O ano de contratação foi representado por variáveis indicadoras. A distribuição regional é desigual: a região Sul concentra cerca de 77% das observações, seguida por Sudeste e Centro-Oeste, enquanto Norte e Nordeste reúnem menos de 1% dos registros, ainda que com severidade média mais alta.

Esclarece-se que a codificação *one-hot* (*dummies*) foi utilizada exclusivamente para a variável ano de contratação (ANO_APOLICE), enquanto todas as demais variáveis categóricas (incluindo município, unidade da federação/região, classificação do produto, cultura e evento preponderante) foram codificadas por *target encoding* em todos os modelos, inclusive no MLG Gama e na RLM. Não foi aplicada codificação *one-hot* às categóricas de alta cardinalidade, dado o risco de explosão dimensional já mencionado.

A multicolinariedade foi investigada no pré-processamento, por meio da matriz de correlação e do Fator de Inflação da Variância (FIV/VIF), tendo motivado a exclusão de variáveis redundantes (ex.: VL_PREMIO_LIQUIDO, VL_SUBVENCAO_FEDERAL e NR_PRODUTIVIDADE_ESTIMADA, que apresentavam VIF > 100 e correlação > 0,99 com outras variáveis). A Tabela 3 reporta o FIV final de cada variável do modelo, calculado sobre a base já depurada.

Tabela 3 - Fator de Inflação da Variância (FIV/VIF) das variáveis do modelo final.

Variável	FIV
Constante	-

Área total	4.60
Número de animais	1.18
Produtividade segurada	1.56
Nível de cobertura	3.12
Limite de garantia	3.58
Município	1.73
Região	1.38
Classificação do produto	1.04
Cultura	1.85
Evento preponderante	2.25
Ano 2017	1.59
Ano 2018	3.10
Ano 2019	2.66
Ano 2020	3.71
Ano 2021	7.07
Ano 2022	2.85
Ano 2023	2.59
Ano 2024	2.59

Fonte: Elaboração própria (2026).

Com exceção da constante (cuja leitura de FIV não é substantiva), todas as variáveis apresentaram FIV inferior a 10 (máximo de 7,07 para o indicador do ano de 2021), o que confirma que a multicolinariedade identificada na etapa de pré-processamento foi adequadamente tratada antes da estimação dos modelos finais. Cabe uma ressalva: o resumo estatístico da RLM (Seção 4.3) reporta ainda um Número de Condição elevado ($3{,}68 \times 10^6$), com alerta automático do software para possível multicolinearidade. Diferentemente do FIV, o número de condição é sensível também à diferença de escala entre a constante e os regressores, podendo permanecer alto mesmo quando o FIV de cada variável, individualmente, é baixo (Belsley; Kuh; Welsch, 1980); a análise por FIV aqui apresentada é, portanto, o critério mais direto para avaliar a multicolinariedade entre os preditores do modelo.

4.2 SELEÇÃO E AVALIAÇÃO DOS MODELOS

Esta seção apresenta os modelos selecionados, seus hiperparâmetros e o desempenho obtido na validação cruzada. Foram avaliados dois modelos estatísticos, a Regressão Linear Múltipla (RLM) e o Modelo Linear Generalizado (MLG) com distribuição Gama e ligação logarítmica, e dois modelos de aprendizado de máquina, o Random Forest e o XGBoost. Os hiperparâmetros dos modelos de aprendizado de máquina foram definidos por busca em grade, adotando o RMSE como critério de seleção; a comparação final entre todos os modelos considerou conjuntamente o MAE e o RMSE, além do sMAPE, de forma a contemplar o erro médio, a penalização de erros elevados e o erro relativo (James *et al.*, 2023).

Antes da modelagem definitiva, verificou-se se a separação da base por tipo de atividade, agrícola e pecuária, traria ganho em relação à modelagem conjunta. Os quatro modelos foram reestimados isoladamente para cada atividade e comparados aos resultados obtidos sobre a base completa. A Tabela 4 apresenta os erros de teste, em reais, nas três situações.

Tabela 4 - Erro de teste na base conjunta e nas bases separadas por atividade.

Modelo	MAE conj.	MAE agríc.	MAE pec.	RMSE conj.	RMSE agríc.	RMSE pec.
GLM Gama	54.528	55.824	38.918	94.377	100.012	70.473
RLM	55.044	55.335	**36.130**	99.997	100.610	67.406
Random Forest	**48.520**	48.031	37.954	83.806	**83.443**	**64.512**
XGBoost	48.850	**48.026**	38.934	**83.628**	83.556	67.167

Fonte: Elaboração própria (2026).

Notas: conj = Base Conjunta; agríc = Base Agrícola; pec = Base Pecuária.

Os três conjuntos são muito desiguais em tamanho: a base conjunta (conj.) reúne 190.423 apólices, das quais 189.849 (99,7%) são agrícolas e apenas 574 (0,3%) pecuárias. Por isso, o erro da base agrícola (agríc.) reproduz quase exatamente o da base conjunta, que é dominada por essas apólices, mantendo-se nesse recorte a vantagem dos algoritmos de árvore sobre os modelos estatísticos. Na base pecuária (pec.), por ser um subconjunto reduzido, homogêneo e associado a perdas de menor magnitude, os erros absolutos são mais baixos, mas o ordenamento entre os modelos se altera: o Random Forest apresenta o menor RMSE (R$ 64.512), ao passo que a RLM registra o menor MAE (R$ 36.130), superando tanto o Random Forest (R$ 37.954) quanto o XGBoost (R$ 38.934), que também perde para a RLM em RMSE quando se considera a proximidade dos valores (R$ 67.167 ante R$ 67.406).

Esse resultado é consistente com a natureza dos algoritmos de árvore: por dependerem de um número maior de parâmetros ajustados iterativamente, tendem a exigir mais dados de treino para estabilizar suas estimativas do que modelos mais parcimoniosos, tornando-se mais sensíveis à variância amostral em subconjuntos pequenos como o de 574 observações pecuárias (Kuhn; Johnson, 2013; James *et al.*, 2023). Como a segmentação não traz ganho preditivo consistente e a base é essencialmente agrícola, adotou-se a base conjunta na modelagem descrita a seguir.

A validação cruzada empregou a estratégia KFold com partições embaralhadas e semente aleatória fixa: dez partições para a RLM e cinco para o MLG Gama, o Random Forest e o XGBoost, número reduzido nesses três casos em razão do custo computacional do reajuste

completo do pré-processamento em cada partição. Em cada partição, o pré-processamento (transformação logarítmica, padronização e target encoding) foi ajustado apenas sobre os dados de treino e aplicado ao conjunto de teste, o que evita o vazamento de informação. A Tabela 5 reúne a especificação de cada modelo e os hiperparâmetros utilizados.

Tabela 5 – Modelos e hiperparâmetros adotados na validação cruzada.

Modelo	Configuração adotada
RLM	OLS; log(1+y); padronização; *target encoding*
MLG Gama	Gama; ligação log; padronização; *target encoding;*
Random Forest	200 árvores; profundidade 18; mín. divisão 10; mín. folha 5; sqrt
XGBoost	200 árvores; taxa 0,05; profundidade 8; subsample 0,8; colsample 0,8

Fonte: Elaboração própria (2026).

Notas: Árvores = quantidade de árvores do modelo; profundidade = tamanho máximo de cada árvore; mín. divisão/folha = número mínimo de observações exigido para o modelo dividir os dados; *max_features, subsample* e *colsample* = proporção de variáveis ou observações sorteadas para treinar cada árvore; taxa = taxa de aprendizado do XGBoost.

Os hiperparâmetros dos modelos de árvore foram definidos por ajuste manual orientado pelo comportamento do erro em validação cruzada, e não pelo resultado de uma busca em grade completa: as execuções de GridSearchCV sobre a base integral não puderam ser concluídas dentro dos limites de tempo e memória do ambiente utilizado. Adotaram-se, assim, configurações regularizadas e conservadoras (Tabela 5), com profundidade limitada, número mínimo de observações por folha e amostragem aleatória de variáveis e observações, escolhas voltadas a conter o sobreajuste característico de modelos de árvore em bases com forte assimetria (Kuhn; Johnson, 2013).

Na validação cruzada, os modelos de aprendizado de máquina apresentaram o menor erro médio. O XGBoost e o Random Forest alcançaram RMSE médio em torno de R$ 83 mil, ante cerca de R$ 94 mil do MLG Gama e R$ 100 mil da RLM. A redução relativa do RMSE em relação ao MLG Gama, tomado como referência estatística, foi de aproximadamente 12%, o que sugere a presença de relações não lineares e de interações entre variáveis produtivas, regionais e climáticas que os modelos lineares não capturam. A diferença entre o XGBoost e o Random Forest foi pequena, indício de que ambos os algoritmos baseados em árvores se ajustam de forma semelhante à estrutura dos dados.

4.3 DESEMPENHO NOS CONJUNTOS DE TREINAMENTO E TESTE

A Tabela 6 apresenta o desempenho dos modelos nos conjuntos de treinamento e teste, em valores de severidade (R$), segundo as métricas MAE, RMSE e sMAPE.

Tabela 6 – Desempenho dos modelos nos conjuntos de treinamento e teste.

Modelo	MAE (treino)	RMSE (treino)	sMAPE (treino)	MAE (teste)	RMSE (teste)	sMAPE (teste)
RLM	55.039	99.994	70,67%	55.044	99.997	70,68%
MLG Gama	54.520	94.370	68,89%	54.528	94.377	68,90%
Random Forest	42.575	72.188	58,82%	48.097	83.445	**62,12%**
XGBoost	43.758	72.348	60,82%	**48.017**	**83.422**	62,49%

Fonte: Elaboração própria (2026).

No conjunto de teste, o XGBoost obteve o menor erro absoluto, com RMSE de R$ 83.422 e MAE de R$ 48.017. Em relação ao MLG Gama, a redução relativa foi de 11,6% no RMSE e de 11,9% no MAE, o que confirma o ganho de desempenho fora da amostra. O Random Forest apresentou métricas muito próximas às do XGBoost (RMSE de R$ 83.445 e MAE de R$ 48.097) e obteve o menor sMAPE (62,12% contra 62,49%). Essa inversão é coerente com a natureza das métricas: o sMAPE é um erro relativo, sensível aos sinistros de menor valor, em que o Random Forest, por produzir previsões mais suaves, comete erros percentuais menores; já o MAE, erro absoluto, e o RMSE, erro quadrático, são dominados pelos sinistros de alto valor, que o XGBoost ajusta com maior precisão. A RLM e o MLG Gama ficaram atrás de ambos, com RMSE de R$ 99.997 e R$ 94.377, respectivamente.

A diferença entre treino e teste deve ser lida em termos comparativos. A RLM e o MLG Gama exibiram erros de treino e teste praticamente idênticos, comportamento coerente com modelos de menor complexidade e menor variância (Montgomery; Peck; Vining, 2012). Os modelos de árvore apresentaram elevação mais acentuada do erro de teste em relação ao de treino: o Random Forest passou de R$ 72.188 para R$ 83.445 e o XGBoost, de R$ 72.348 para R$ 83.422, o que indica não um sobreajuste em termos absolutos, mas um sobreajustamento relativamente maior do que o dos modelos estatísticos. Esse afastamento foi contido pela regularização adotada em ambos: profundidade limitada, número mínimo de amostras por folha e amostragem aleatória de variáveis no Random Forest, e subamostragem de observações e variáveis, com taxa de aprendizado reduzida, no XGBoost.

A comparação entre as duas abordagens pode ser organizada em três dimensões. Quanto ao desempenho preditivo, os modelos de aprendizado de máquina superaram os estatísticos, sobretudo no erro absoluto fora da amostra. Quanto à estabilidade, os modelos estatísticos mostraram-se mais regulares entre treino e teste, enquanto os de árvore, embora mais precisos, apresentaram maior sobreajustamento relativo. Quanto à interpretabilidade, os modelos estatísticos permitem a leitura direta dos efeitos por meio dos coeficientes, ao passo que os de árvore exigem técnicas auxiliares, como o SHAP. Esse compromisso entre acurácia e transparência está alinhado a Mahajan *et al.* (2025), que observaram maior acurácia dos métodos de aprendizado de máquina acompanhada de menor transparência frente aos modelos estatísticos.

A análise dos resíduos da RLM reforça a natureza da variável resposta. O diagnóstico do modelo apresentou estatística de Durbin-Watson de 1,73, assimetria de −1,47 e curtose de 8,28, valores que indicam resíduos não normais, com forte concentração e cauda longa. Esse afastamento da normalidade é esperado em dados de severidade agrícola e reforça a adequação de modelos flexíveis, menos dependentes da hipótese de erros normais (Montgomery; Peck; Vining, 2012).

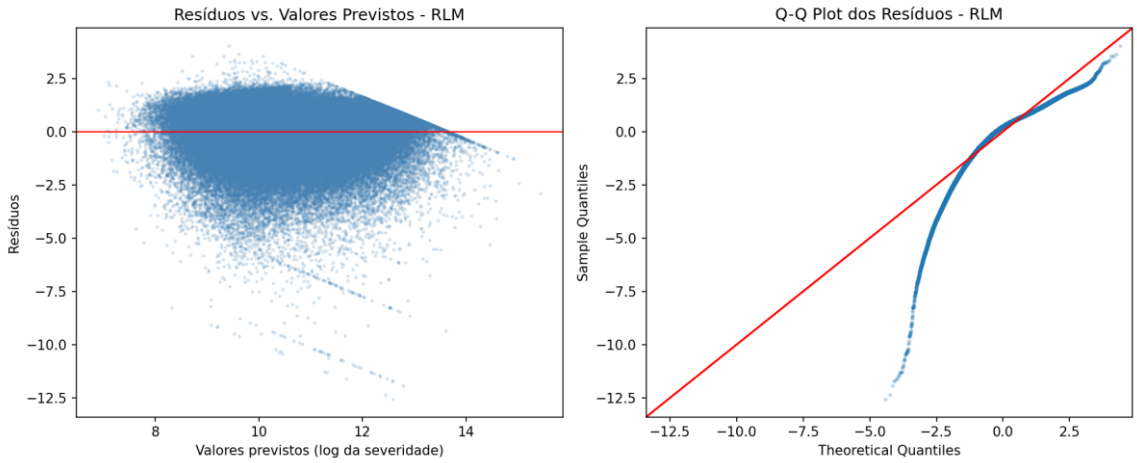
O resumo estatístico da RLM reporta também o Número de Condição (Cond. No.), estatística apresentada no mesmo bloco de saída do Durbin-Watson, cujo valor ($3{,}68 \times 10^6$) vem acompanhado do alerta automático do software de que isso pode indicar multicolinearidade ou outros problemas numéricos. Esse número de condição elevado é consistente com a preocupação de multicolinearidade já tratada na Seção 4.1; contudo, como o FIV de cada variável do modelo final é inferior a 10 (Tabela 3), a multicolinearidade entre os regressores não explica, isoladamente, um número de condição tão alto.

Esse indicador é também sensível à diferença de escala entre a constante (intercepto) e os regressores padronizados/*target-encoded*, um problema numérico distinto da multicolinearidade genuína entre variáveis explicativas (Belsley; Kuh; Welsch, 1980), e não invalida as conclusões baseadas no FIV. Ainda assim, o achado reforça a cautela já recomendada na leitura dos erros-padrão e das inferências da RLM.

Complementarmente à análise de resíduos (Figura 10), aplicou-se o teste de Breusch-Pagan sobre os resíduos da RLM, que avalia a heterocedasticidade, isto é, a dependência da variância do erro em relação às variáveis explicativas. O teste rejeitou a hipótese de homocedasticidade (estatística LM = 5.291,9; p-valor < 0,001), resultado coerente com a correlação dos resíduos ao quadrado reportada na Seção 4.1, que apontou o nível de cobertura, o evento preponderante e o ano de 2021 como principais fontes de variação. Na prática, isso

significa que os erros-padrão e os testes de significância da RLM devem ser lidos com cautela, pois a violação da homocedasticidade os torna inconsistentes (Wooldridge, 2016). O MLG Gama mitiga parcialmente esse problema, pois assume variância proporcional ao quadrado da média ($\text{Var}(Y) \propto \mu^2$), estrutura mais adequada a dados de severidade do que a variância constante pressuposta pela RLM (Montgomery; Peck; Vining, 2012).

Figura 10 – Resíduos vs. valores previstos e Q-Q plot dos resíduos (RLM).



Fonte: Elaboração própria (2026).

O gráfico de resíduos contra os valores previstos (Figura 10, painel esquerdo) evidencia um leque que se estreita à medida que o valor previsto aumenta, assinatura visual da heterocedasticidade confirmada pelo teste de Breusch-Pagan, além de uma faixa diagonal de observações subestimadas nos valores mais altos, compatível com o truncamento da variável resposta no percentil 99 adotado no pré-processamento. O Q-Q plot (painel direito) mostra forte afastamento da normalidade na cauda esquerda, refletindo a assimetria negativa (-1,76) e a curtose elevada (9,73) já reportadas.

Em conjunto, os três diagnósticos (Durbin-Watson, Breusch-Pagan e Q-Q plot) convergem para uma mesma conclusão substantiva do trabalho: a RLM, apoiada nas hipóteses clássicas de erros normais e homocedásticos, está estatisticamente mal especificada para a severidade do seguro agropecuário. Isso explica, em bases estatísticas e não apenas preditivas, por que o MLG Gama, que relaxa a normalidade e ajusta a variância à média, e os modelos de árvore, que dispensam qualquer hipótese distribucional, apresentaram desempenho superior nas Tabelas 4 e 6. Os diagnósticos de resíduos reforçam, portanto, que a escolha do modelo não é apenas uma questão de acurácia, mas de adequação da estrutura probabilística aos dados de severidade.

4.4 EXPLICABILIDADE DOS MODELOS

A interpretabilidade foi avaliada de duas formas: pela leitura direta dos coeficientes nos modelos estatísticos e pela aplicação do SHAP no XGBoost.

Tabela 7 – Coeficientes do MLG Gama na escala multiplicativa.

Variável	Coef. (β)	exp(β)	Efeito	p-valor
Intercepto	9,986	21.712,23	—	< 0,001
Limite de garantia	0,517	1,676	+67,6%	< 0,001
Área total	0,257	1,293	+29,3%	< 0,001
Produtividade segurada	0,129	1,138	+13,8%	< 0,001
Nível de cobertura	-0,041	0,960	−4,0%	< 0,001
Número de animais	-0,021	0,979	−2,1%	< 0,001
Ano de 2021	0,449	1,567	+56,7%	< 0,001

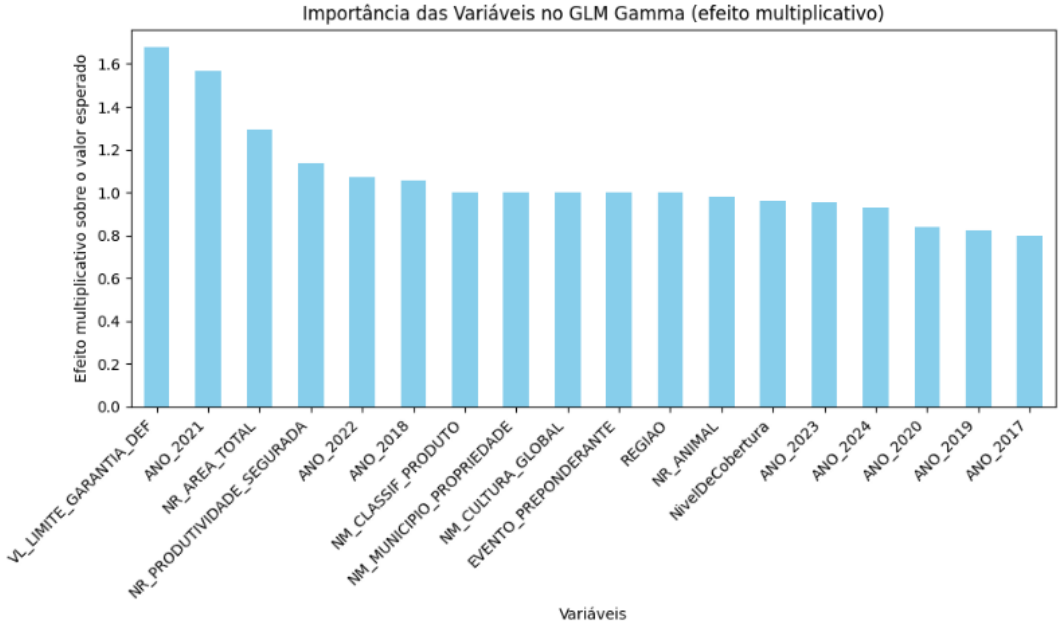
Fonte: Elaboração própria (2026).

A Tabela 7 apresenta os coeficientes do MLG Gama na escala multiplicativa, exp(β). Todos os preditores numéricos foram estatisticamente significativos ($p < 0,001$). O limite de garantia exerceu o maior efeito: o aumento de um desvio-padrão eleva a severidade esperada em cerca de 68%, resultado coerente com a lógica atuarial, uma vez que a exposição financeira do contrato condiciona diretamente a perda esperada (Denuit; Hainaut; Trufin, 2019). A área total (+29%) e a produtividade segurada (+14%) também elevaram a severidade, enquanto o nível de cobertura e o número de animais apresentaram efeito de redução, de menor magnitude.

Destaca-se que "estatisticamente significativo" ($p < 0,001$) indica apenas que o coeficiente é distinto de zero com alto grau de confiança, dado o tamanho amostral elevado ($n = 190.423$), o que facilita a obtenção de significância mesmo para efeitos pequenos, e não deve ser interpretado como sinônimo de maior explicabilidade ou importância prática da variável. A magnitude do efeito (exp(β)) e a explicabilidade relativa de cada variável são avaliadas separadamente, pela ordenação dos efeitos multiplicativos (Figura 11) e, para os modelos de árvore, pelos valores de SHAP (Seção 4.4.1).

O ano de 2021 destacou-se com o maior efeito positivo, compatível com a crise hídrica e com as fortes geadas registradas naquele ano, que atingiram sobretudo o Centro-Sul do país, principal região produtora da base, e levaram o Programa de Subvenção ao Prêmio do Seguro Rural ao maior volume de indenizações de sua série histórica (Brasil, 2022b). O pseudo-R^2 de 0,70 indica bom ajuste do modelo.

Figura 11 – Importância das variáveis no MLG Gama (efeito multiplicativo).



Fonte: Elaboração própria (2026).

A Figura 11 ordena as variáveis pelo efeito multiplicativo estimado no MLG Gama. O limite de garantia e o ano de 2021 aparecem como os fatores de maior impacto sobre a severidade esperada, seguidos pela área total, o que traduz visualmente as magnitudes da Tabela 7. Já a Tabela 8 trás os coeficientes da RLM.

Tabela 8 – Coeficientes da RLM (variável dependente em logaritmo).

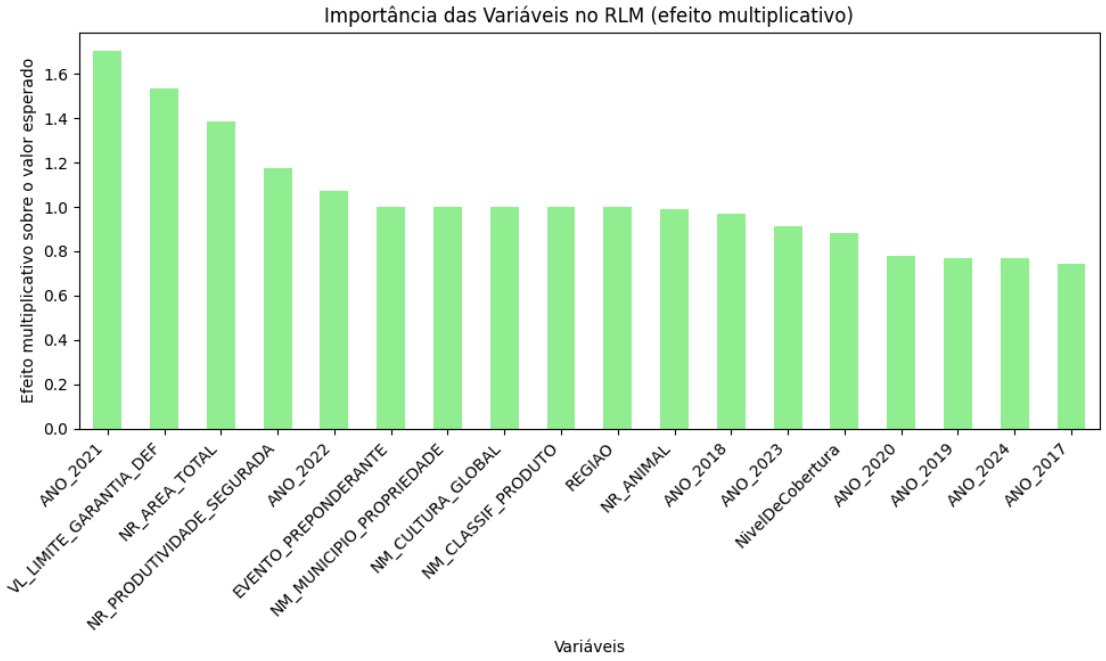
Variável	Coef.	exp(coef)	p-valor
Intercepto	9,416	12.285,25	< 0,001
Limite de garantia	0,429	1,535	< 0,001
Área total	0,327	1,386	< 0,001
Produtividade segurada	0,161	1,174	< 0,001
Nível de cobertura	-0,125	0,883	< 0,001
Número de animais	-0,010	0,990	< 0,001

Fonte: Elaboração própria (2026).

A RLM, de acordo com a Tabela 8, com a variável dependente em logaritmo, conduziu às mesmas direções de efeito observadas no MLG Gama (Tabela 8), com destaque novamente para o limite de garantia e a área total. O coeficiente de determinação ($R^2 = 0,42$) indica que o modelo linear explica menos da metade da variabilidade da severidade. A comparação com o MLG Gama é instrutiva: embora ambos concordem na direção dos efeitos, o MLG Gama ajusta-

se melhor à distribuição da severidade que geralmente é positiva e assimétrica, com pseudo-R^2 de 0,70, ao passo que a RLM pressupõe erros normais, hipótese não atendida pelos dados (assimetria de −1,47 e curtose de 8,28 nos resíduos). Assim, o MLG Gama é estatisticamente mais adequado à natureza da variável resposta, enquanto a RLM oferece uma leitura mais simples, porém menos aderente à cauda pesada das indenizações.

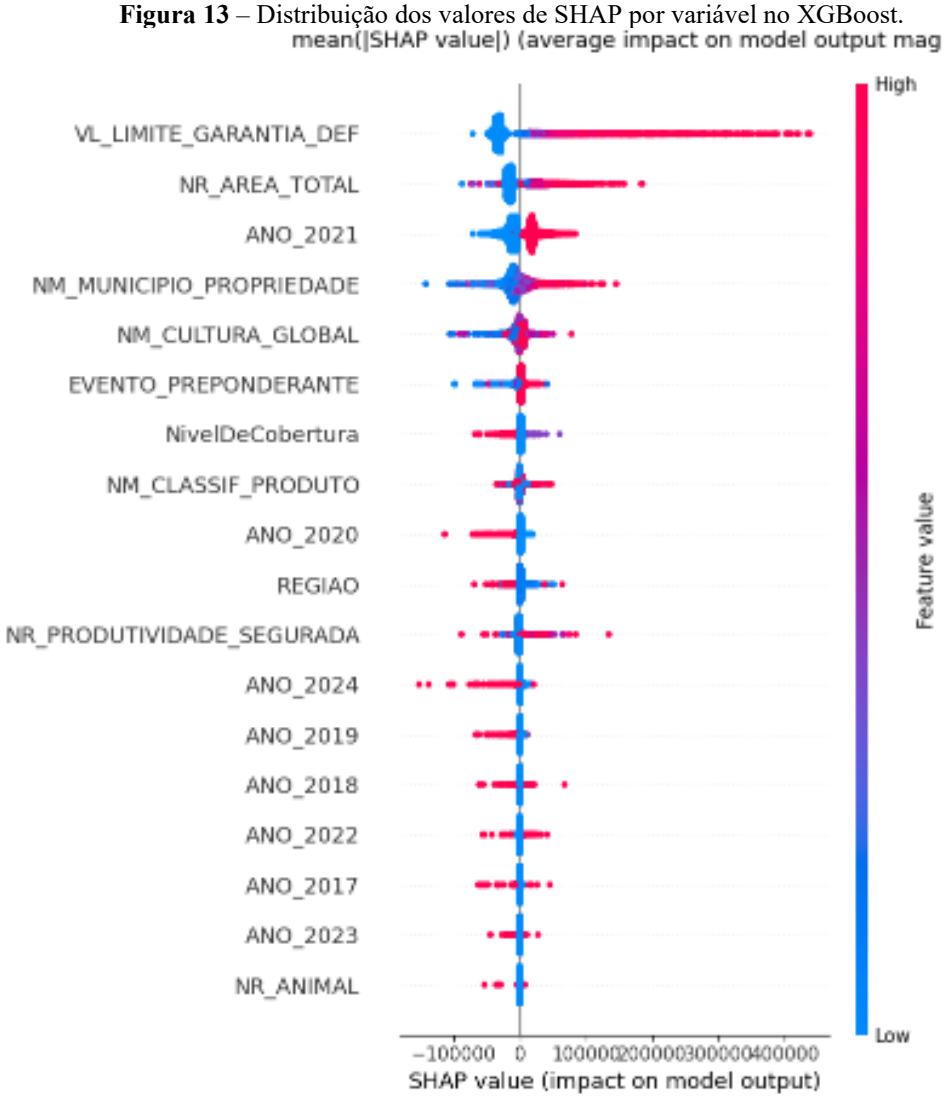
Figura 12 – Importância das variáveis na RLM (efeito multiplicativo).



Fonte: Elaboração própria (2026).

A Figura 12 ordena as variáveis pelo efeito multiplicativo estimado na RLM. O ordenamento é praticamente o mesmo observado no MLG Gama (Figura 11), com inversão apenas entre as duas primeiras posições: na RLM, o ano de 2021 assume o maior efeito (1,70), seguido do limite de garantia (1,54), ao passo que no MLG Gama o limite de garantia lidera (1,68), seguido do ano de 2021 (1,57). A convergência entre as duas especificações reforça a estabilidade dos determinantes identificados.

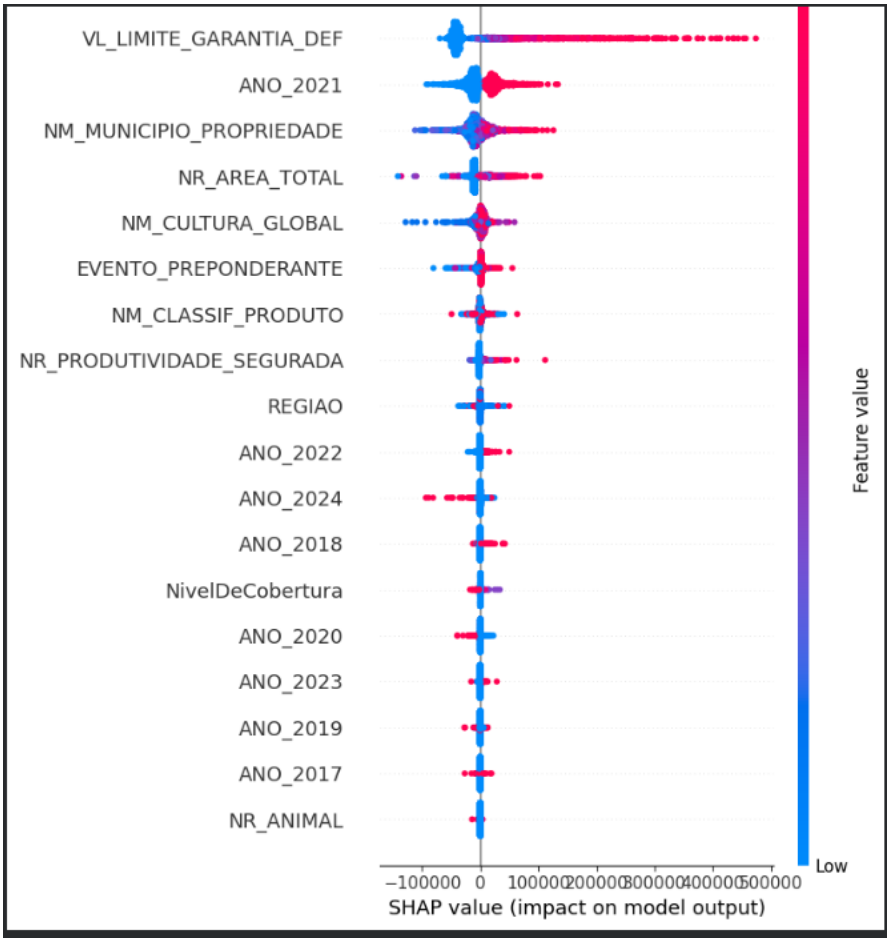
Figura 13 – Distribuição dos valores de SHAP por variável no XGBoost.



Fonte: Elaboração própria (2026).

A Figura 13 apresenta os valores de SHAP do XGBoost, que decompõem a contribuição de cada variável para a previsão. O limite de garantia foi, com folga, a variável mais influente, seguido pela área total, pelo ano de 2021, pelo município e pela cultura; os pontos à direita, associados a valores altos dessas variáveis, elevam a severidade prevista. Esse ordenamento confirma as conclusões dos modelos estatísticos: a exposição financeira do contrato e a dimensão da área são os principais determinantes da severidade, enquanto a localização, captada pelo município e pela região, e a cultura segurada exercem influência relevante, porém secundária. A convergência entre os efeitos do MLG e o ordenamento do SHAP reforça a plausibilidade dos resultados no contexto agrícola, em que o valor segurado, a escala da produção e os fatores climáticos e geográficos condicionam as perdas indenizáveis (Cartolano *et al.*, 2024).

Figura 14 – Distribuição dos valores de SHAP por variável no Random Forest.



Fonte: Elaboração própria (2026).

A Figura 14 apresenta os valores de SHAP para o Random Forest. O ordenamento das variáveis é praticamente o mesmo obtido para o XGBoost e para o MLG Gama: o limite de garantia domina, seguido do ano de 2021, do município, da área total e da cultura. Essa convergência entre os três modelos, dois de árvore e um estatístico, reforça a robustez dos resultados, indicando que a exposição financeira do contrato e os fatores climáticos e geográficos são os determinantes centrais da severidade, independentemente da classe de modelo adotada.

Os resultados conversam com a literatura. O melhor desempenho dos modelos de árvore na previsão da severidade é compatível com Su e Bai (2020), que verificaram ganho do XGBoost sobre os MLGs ao capturar interações e dependências não lineares, e com Noorunnahar *et al.* (2023), que reportaram baixo erro do XGBoost em aplicações agrícolas. A relevância das variáveis de exposição e de localização também dialoga com Brati *et al.* (2025), que identificaram nos modelos de árvore boa capacidade de ordenar sinistros de alto custo. Por

outro lado, o resultado diverge parcialmente de Clemente *et al.* (2023), para quem os MLGs superaram o aprendizado de máquina na modelagem da severidade, diferença que pode ser atribuída à natureza climática e à heterogeneidade regional dos dados de seguro rural. Por fim, o uso do SHAP, na linha de Zhang (2024) e Lundberg *et al.* (2023), tornou o modelo de melhor desempenho interpretável, atendendo à exigência de transparência do setor segurador sem comprometer a acurácia preditiva.

Em síntese, o XGBoost apresentou o melhor equilíbrio entre desempenho preditivo e estabilidade, com erro fora da amostra inferior ao dos modelos estatísticos e afastamento controlado entre treino e teste. O Random Forest alcançou resultado praticamente equivalente, com vantagem no sMAPE, métrica que pondera mais fortemente os sinistros de menor valor. Os modelos estatísticos, embora menos precisos, preservaram a vantagem da interpretação direta dos coeficientes. A combinação do XGBoost com o SHAP atende ao objetivo do estudo ao oferecer um modelo simultaneamente preciso e explicável para a previsão da severidade no seguro agropecuário.

4.5 ANÁLISE DE ROBUSTEZ PARA MODELOS SEM O LIMITE DE GARANTIA

Como o limite de garantia representa a exposição financeira máxima assumida pela seguradora no contrato, sua presença entre os preditores levanta a questão de quanto do desempenho dos modelos decorre especificamente dessa variável. Para avaliar essa dependência, os quatro modelos foram reestimados excluindo o limite de garantia do conjunto de preditores, mantendo-se as demais variáveis, os mesmos hiperparâmetros da Tabela 5, o mesmo esquema de validação cruzada (KFold) e o mesmo pré-processamento ajustado apenas sobre os dados de treino. A comparação é apresentada na Tabela 9, com base nos erros de teste.

Tabela 9 – Erros de teste dos modelos com e sem o limite de garantia como preditor.

Modelo	MAE (com)	MAE (sem)	Δ MAE	RMSE (com)	RMSE (sem)	Δ RMSE	sMAPE (com)	sMAPE (sem)
RLM	55.044	57.003	**+3,6%**	99.997	108.275	+8,3%	70,68%	71,11%
MLG Gama	54.528	57.900	+6,2%	94.377	104.932	+11,2%	68,90%	70,58%
Random Forest	48.097	51.564	+7,2%	83.445	89.010	+6,7%	**62,12%**	65,27%
XGBoost	**48.017**	**51.061**	+6,3%	**83.422**	**88.432**	**+6,0%**	62,49%	**64,95%**

Fonte: Elaboração própria (2026).

A remoção do limite de garantia deteriora o desempenho de todos os modelos, mas o padrão da perda difere conforme a métrica considerada. No RMSE, métrica mais sensível aos sinistros de alto valor, os modelos de árvore mostram-se claramente mais robustos: o XGBoost perde 6,0% e o Random Forest, 6,7%, ante 11,2% do MLG Gama e 8,3% da RLM. No MAE, contudo, a RLM apresenta a menor deterioração do conjunto (3,6%), seguida do MLG Gama (6,2%), do XGBoost (6,3%) e do Random Forest (7,2%).

A leitura conjunta sugere que os algoritmos de árvore recuperam, a partir das demais variáveis, sobretudo a informação necessária para prever os sinistros de maior magnitude, ao passo que a perda no erro médio se distribui de forma mais uniforme entre as abordagens. Esse comportamento é coerente com a caracterização de Faceli *et al.* (2025) sobre a capacidade dos métodos baseados em árvores de representar relações não lineares e interações entre atributos sem especificação prévia, e com a observação de James *et al.* (2023) de que modelos lineares dependem mais fortemente da presença explícita de cada variável relevante, por não reconstruírem efeitos combinados de forma automática.

A redistribuição do efeito entre as variáveis explica esse resultado. Nos coeficientes do MLG Gama, a área total passa de 0,257 (efeito multiplicativo de 1,29) na especificação completa para 0,662 (efeito de 1,94) na especificação sem o limite de garantia, e a produtividade segurada passa de 0,129 para 0,274, absorvendo parte substancial do efeito antes atribuído ao limite de garantia (0,517). Esse deslocamento é esperado diante da correlação de 0,71 entre área total e limite de garantia observada na Figura 8: quando um preditor correlacionado é omitido, seu efeito é parcialmente incorporado pelos remanescentes, fenômeno discutido por Gujarati e Porter (2011) ao tratar das consequências da omissão de variáveis relevantes em modelos de regressão.

Em termos de ajuste, o pseudo-R^2 (Cox-Snell) do MLG Gama cai de 0,70 para 0,49 e o R^2 da RLM recua de 0,421 para 0,398, confirmando que o limite de garantia concentra parcela expressiva da explicação intra-amostral dos modelos estatísticos. A queda no ajuste, contudo, é proporcionalmente maior que a perda de acurácia preditiva fora da amostra, o que indica que parte do poder explicativo dessa variável é redundante com características já capturadas pelo porte da apólice. Em conjunto, os resultados sustentam a decisão de manter o limite de garantia na especificação principal, por ser a medida direta da exposição financeira do contrato e, portanto, interpretativamente relevante para a precificação atuarial (Denuit; Hainaut; Trufin, 2019).

Cabe destacar, ainda, que a vantagem dos modelos de aprendizado de máquina não se restringe à acurácia, estendendo-se à robustez frente a alterações na especificação do conjunto de preditores, particularmente na previsão dos sinistros de maior magnitude, atributo especialmente útil em aplicações securitárias, nas quais a disponibilidade de variáveis contratuais pode variar entre carteiras e períodos (Aas *et al.*, 2024). Ainda assim, a perda observada em todos os modelos confirma que o limite de garantia não é redundante: nenhuma das abordagens reconstrói integralmente a informação que ele carrega.

5 CONSIDERAÇÕES FINAIS

A previsão da severidade dos sinistros do seguro agropecuário, com base nas apólices do Programa de Subvenção ao Prêmio do Seguro Rural recepcionadas entre 2016 e 2024, colocou frente a frente os modelos estatísticos tradicionais e os modelos interpretáveis de aprendizado de máquina. Do confronto entre as duas abordagens emergem diferenças nítidas de desempenho e de interpretabilidade.

Aplicados sobre a mesma base de severidade, com pré-processamento e validação cruzada padronizados, os modelos de aprendizado de máquina apresentaram desempenho preditivo superior ao dos estatísticos. O XGBoost alcançou o menor erro fora da amostra, com redução de cerca de 12% no RMSE em relação ao modelo linear generalizado, e o Random Forest obteve resultado praticamente equivalente. Os modelos estatísticos, ainda que menos precisos, mostraram-se mais estáveis entre treino e teste, enquanto os modelos de árvore apresentaram sobreajuste um pouco maior, contido pela regularização. Esse comportamento revela que a severidade dos sinistros agropecuários envolve relações não lineares e interações pouco captadas por especificações lineares.

A análise de robustez conduzida na Seção 4.5, que reestimou os quatro modelos sem o limite de garantia, evidenciou ainda uma vantagem adicional dos algoritmos baseados em árvores: enquanto o MLG Gama perdeu 11,2% de acurácia em RMSE com a retirada dessa variável e a RLM, 8,3%, o XGBoost limitou a perda a 6,0% e o Random Forest, a 6,7%, recuperando parte relevante da informação omitida a partir das demais características da apólice.

A vantagem preditiva, contudo, não implicou perda de transparência. Os coeficientes dos modelos estatísticos permitiram a leitura direta da direção e da magnitude dos efeitos, e a técnica SHAP tornou o XGBoost igualmente interpretável. As explicações convergiram ao apontar o limite de garantia e o ano de 2021 como os principais determinantes da severidade, seguidos da área total, o que confere consistência aos resultados e plausibilidade no contexto agrícola.

A principal contribuição do trabalho está em demonstrar que é possível conciliar precisão e transparência na modelagem do risco agropecuário. Do ponto de vista atuarial, os resultados oferecem subsídios para a precificação e a mitigação do risco no seguro rural; do ponto de vista metodológico, evidencia-se que a combinação de um modelo de aprendizado de máquina com técnicas de explicabilidade pode substituir, com ganho de acurácia e sem perda de interpretabilidade, o uso isolado de modelos estatísticos tradicionais.

O estudo apresenta limitações que devem ser consideradas. A exposição financeira do contrato foi representada pelo limite de garantia, e não pelo prêmio, o que pode captar apenas parte do risco; a variável de severidade foi truncada no percentil 99, o que atenua a influência dos sinistros extremos, justamente os de maior relevância atuarial; e a validação cruzada foi aleatória, sem considerar a estrutura temporal dos dados, o que pode superestimar o desempenho fora da amostra em um cenário de previsão de anos futuros.

Sugere-se para trabalhos futuros a modelagem separada da cauda de perdas extremas, a adoção de validação temporal, a incorporação de variáveis climáticas e de zoneamento de risco e a extensão da análise à frequência de sinistros, além da severidade. Recomenda-se também investigar especificações alternativas da variável resposta, como a razão entre a indenização e o limite de garantia, que representaria a severidade em termos relativos à exposição contratada. Ainda assim, os resultados reforçam o potencial dos modelos de aprendizado de máquina interpretáveis como ferramenta de apoio às seguradoras, ao aliar maior precisão preditiva à transparência exigida pelo setor na gestão do risco agropecuário.

REFERÊNCIAS

AAS, Kjersti; CHARPENTIER, Arthur; HUANG, Fei; RICHMAN, Ronald. **Insurance analytics:** prediction, explainability, and fairness. Annals of Actuarial Science, v. 18, p. 535–539, 2024. Disponível em: <https://doi.org/10.1017/S1748499524000289>. Acesso em: 28 jun. 2025.

AGROPECUÁRIA. *In.*: DICIO, Dicionário Online de Português. Porto: 7Graus, 2024. Disponível em: https://www.dicio.com.br/agropecuaria/. Acesso em: 24 jul. 2024.

AJZAN, Salame Zayd; BABAMURATOV, Bekzod; AYYAPPAN, V.; PRABAKARAN, N. **Modelling crop yield prediction with random forest and remote sensing data**. ResearchGate, 2023. Disponível em: <https://www.researchgate.net/publication/395630727_Modelling_Crop_Yield_Prediction_with_Random_Forest_and_Remote_Sensing_Data>. Acesso em: 26 jul. 2026.

ALMEIDA, António Betâmio de. **Gestão do risco e da incerteza:** conceitos e filosofia subjacente. In: LOURENÇO, Luciano (org.). Realidades e desafios na gestão dos riscos: diálogo entre ciência e utilizadores. Coimbra: Imprensa da Universidade de Coimbra, 2023. p. 19–29. Disponível em: https://www.uc.pt/fluc/nicif/Publicacoes/livros/dialogos/Artg02.pdf. Acesso em: 26 jun. 2025.

ARAÚJO, Sara Oleiro; PERES, Ricardo Silva; RAMALHO, José C.; LIDON, Fernando Cebola. **Machine learning applications in agriculture: current trends, challenges and future perspectives**. ResearchGate, 2023. Disponível em: <https://www.researchgate.net/publication/376146771_Machine_Learning_Applications_in_Agriculture_Current_Trends_Challenges_and_Future_Perspectives>. Acesso em: 26 jul. 2026.

BARBOSA, Vivian Cristina Ribeiro; BRISOLA, Marlon Vinícius. **Além dos campos:** as prospecções tecnológicas sustentáveis da EMBRAPA para o agronegócio brasileiro. Revista de Economia e Sociologia Rural, v. 62, n. 3, e270441, 2024. Disponível em: https://www.scielo.br/j/resr/a/KvCjrVjRPJGJ3T57CC3X9zK/?format=pdf&lang=pt. Acesso em: 6 nov. 2025.

BENOIT, Kenneth. **Linear Regression Models with Logarithmic Transformations.** London School of Economics, Methodology Institute, 2011. Disponível em: https://kenbenoit.net/assets/courses/ME104/logmodels2.pdf. Acesso em: 20 ago. 2025.

BECKER, João L**. Estatística básica** . Porto Alegre: Bookman, 2015. E-book. p.[Inserir número da página]. ISBN 9788582603130. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788582603130/. Acesso em: 26 jul. 2026.

BELSLEY, David A.; KUH, Edwin; WELSCH, Roy E. **Regression Diagnostics:** Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons, 1980.

BIAUGINI, Matteo. **Interpretable deep learning models for agribusiness:** balancing accuracy and transparency. arXiv preprint, 2023. Disponível em: <https://arxiv.org/pdf/2212.03114>. Acesso em: 1 jul. 2025.

BOEHM, Matthias; KUMAR, Arun; YANG, Jun. **Data Management in Machine Learning Systems**. [*S. l.*]: Springer, 2019. v. 1. Disponível em: https://link.springer.com/book/10.1007/978-3-031-01869-5. Acesso em: 28 jul. 2024.

BURNS, Eileen; DAN, Gene; LARSON, Anders; MEYER, Bob; MOTIWALLA, Zohair; YOLLIN, Guy. **Considerations for Predictive Modeling in Insurance Applications.** Society of Actuaries, maio 2019. Disponível em: <https://www.soa.org/globalassets/assets/files/resources/research-report/2019/considerations-predictive-modeling.pdf>. Acesso em: 6 ago. 2026.

BUSSAB, Wilton de Oliveira; MORETTIN, Pedro Alberto. **Estatística Básica.** 6. ed. São Paulo: Saraiva, 2010. ISBN 978-85-02-08177-2.Disponível em: <https://www.professores.uff.br/jutriavaldes/wp-content/uploads/sites/196/2019/08/Book_EstatBas-Morettin-Bussab.pdf>.Acesso em: 5 mar. 2026.

BLEI, David M. **Linear regression, logistic regression, and generalized linear models.** Columbia University, 2015. Disponível em: https://www.cs.columbia.edu/~blei/fogm/2016F/doc/glms.pdf. Acesso em: 20 ago. 2025.

BLOCKEEL, Hendrik; DEVOS, Laurens; FRÉNAY, Benoît; NANFACK, Géraldin; NIJSSEN, Siegfried. **Decision trees:** from efficient prediction to responsible AI. Frontiers in Artificial Intelligence, v. 6, 25 jul. 2023. Disponível em: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1124553/full. Acesso em: 16 ago. 2025.

BRASIL. **Lei nº 15.040, de 27 de dezembro de 2024.** Dispõe sobre o contrato de seguro e revoga dispositivos da Lei nº 10.406, de 10 de janeiro de 2002 (Código Civil). Diário Oficial da União: seção 1, Brasília, DF, 30 dez. 2024. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/L15040.htm. Acesso em: 4 fev. 2026.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento (MAPA). **Agropecuária Brasileira em Números**. Brasília: MAPA, 2024. Disponível em: <https://www.gov.br/agricultura>. Acesso em: 1 jul. 2025.

BRASIL. Ministério da Agricultura e Pecuária. **Valor da produção agropecuária fecha 2022 em R$ 1,189 trilhão**. 2022a. Disponível em: https://www.gov.br/agricultura/pt-br/assuntos/noticias/valor-da-producao-agropecuaria-fecha-2022-em-r-1-189-trilhao. Acesso em: 13 jul. 2025.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. **Histórico de perdas na agricultura brasileira:** 2000-2021. Brasília: MAPA/AECS, 2022b. ISBN 978-85-7991-185-9. 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. Acesso em: 15 jul. 2026.

BRATI, Esmeralda; BRAIMLLARI, Alma; GJEÇI, Ardit. **Machine Learning Applications for Predicting High-Cost Claims Using Insurance Data.** *Data*, v. 10, n. 6, p. 90, 2025. Disponível em: https://www.mdpi.com/2306-5729/10/6/90. Acesso em: 24 ago. 2025.

BREIMAN, Leo**. Random forests.** Machine Learning, v. 45, n. 1, p. 5–32, 2001. Disponível em: https://doi.org/10.1023/A:1010933404324 (doi.org in Bing). Acesso em: 26 jul. 2026.

BREUSCH, T. S.; PAGAN, A. R. **A Simple Test for Heteroscedasticity and Random Coefficient Variation.** Econometrica, v. 47, n. 5, p. 1287-1294, 1979.

CAFFAGNI, Luiz Cláudio; PAIXÃO, Leonardo; RIOS, Márcio. **Principais tipos de seguro agrícola.** Revista Agroanalysis – FGV, São Paulo, v. 42, n. 3, p. 34–36, mar. 2022. Disponível em: https://periodicos.fgv.br/agroanalysis/article/download/88009/82773/192919. Acesso em: 25 jun. 2025.

CARTOLANO, Daniel C.; MOHAN, Jagan R. N. V.; RAYANOOTHALA, Pravallika Sree; SREE, Praneetha R. **Explainable AI for climate-aware crop yield prediction**. Multimedia Tools and Applications, v. 83, 2024. Disponível em: <https://link.springer.com/article/10.1007/s11042-023-17978-z>. Acesso em: 1 jul. 2025.

CEPEA. **PIB do Agronegócio Brasileiro**. Piracicaba: Centro de Estudos Avançados em Economia Aplicada, ESALQ/USP, 2025. Disponível em: https://www.cepea.org.br/br/pib-do-agronegocio-brasileiro.aspx. Acesso em: 6 nov. 2025.

CORDEIRO, Gauss M.; DEMÉTRIO, Clarice G. B. **Modelos Lineares Generalizados e Extensões.** Recife: Universidade Federal de Pernambuco, 2008. Disponível em: https://www2.ufjf.br/clecio_ferreira//files/2013/05/Livro-Gauss-e-Clarice.pdf. Acesso em: 20 ago. 2025.

CHEN, Chen; KIRABAEVA, Kolarai; KOLERUS, Christina; PARRY, Ivan W. H; VERNON, Nate. **Changing Climate in Brazil:** Key Vulnerabilities and Opportunities. IMF Working Paper No. 24/185. Washington, D.C.: International Monetary Fund, 2024. Disponível em: <https://www.elibrary.imf.org/view/journals/001/2024/185/article-A001-en.xml>. Acesso em: 1 jul. 2025.

CLEMENTE, Carina de Miranda; GUERREIRO, Gracinda Rita Diogo; BRAVO, Jorge Miguel. **Gradient Boosting in Motor Insurance Claim Frequency Modelling.** In: CAPSI 2023 Proceedings. 2023. Disponível em: https://aisel.aisnet.org/capsi2023/5/. Acesso em: 24 ago. 2025.

DAL POZZO, Beatriz Salandin; GASPARRETTO, Suelen Cristina; BENICIO, Luana Maria; GOMES, Pâmela; OZAKI, Vitor Augusto. **Contexto histórico e atual do seguro pecuário no Brasil e no mundo**. Revista de Política Agrícola, v. 32, n. 1, p. 1–23, jan./mar. 2023. Disponível em: https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/seguro-rural/observatorio-do-seguro-rural/estudos/estudos-2023/2023-pozzo-gasparetto-benicio-gomes-ozaki-contexto-historico-e-atual-do-seguro-pecuario-no-brasil-e-no-mundo.pdf. Acesso em: 6 nov. 2025.

DENUIT, Michel; HAINAUT, Donatien; TRUFIN, Julien. **Effective Statistical Learning Methods for Actuaries I:** GLMs and Extensions. Cham: Springer Nature Switzerland AG, 2019. (Springer Actuarial Lecture Notes). ISBN 978-3-030-25819-1.

EMBRAPA. **Impactos das mudanças climáticas na agropecuária.** Em: ATER+ Digital – Mudanças Climáticas. 2023. Disponível em: https://www.atermaisdigital.cnptia.embrapa.br/web/mudancas-climaticas/impactos-na-agropecuaria. Acesso em: 26 jun. 2025.

EMBRAPA. **Pesquisa Agropecuária Brasileira (PAB).** Brasília: Embrapa, 2024. Disponível em: <https://apct.sede.embrapa.br/pab/index>. Acesso em: 1 jul. 2025.

EMBRAPA; MAPA. **Visão 2030:** o futuro da agricultura brasileira. Brasília, DF: Embrapa, 2018. Disponível em: https://www.embrapa.br/visao-de-futuro. Acesso em: 4 fev. 2026.

FAO; OPAS; CEPAL. **The Impact of Disasters on Agriculture and Food Security 2025**: Digital solutions for reducing risks and impacts. Rome: FAO, 2025. Disponível em: https://openknowledge.fao.org/items/74d08f97-306a-4653-8ffe-140a2fc4d783. Acesso em: 26 jul. 2026.

FAUZAN, Muhammad Arief; MURFI, Hendri. **The Accuracy of XGBoost for Insurance Claim Prediction.** International Journal of Advance Soft Computing Applications, v. 10, n. 2, p. 159–171, jul. 2018. Disponível em: https://www.i-csrs.org/Volumes/ijasca/11_IJASCA_The-accuracy-of-XGBoost_159-171.pdf. Acesso em: 21 ago. 2025.

FACELI, Katti; LORENA, Ana C.; GAMA, João; ALMEIDA, Tiago Agostinho De; CARVA, André CPL F De. **Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina .** 3.ed. Rio de Janeiro: LTC, 2025. E-book. p.Capa. ISBN 9788521639213. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788521639213/. Acesso em: 26 jul. 2026.

FEIJÓ, Janaína; ZAHAR, Helena. **Mercado de Trabalho em Pauta – Abril de 2025**. Rio de Janeiro: FGV IBRE, 2025. Disponível em: https://ibre.fgv.br/sites/ibre.fgv.br/files/arquivos/u65/relatorio_fev2025_observatorio_v3.pdf. Acesso em: 6 nov. 2025.

FOX, John; WEISBERG, Sanford. **car:** Companion to Applied Regression. Versão 3.1-3. Vienna: R Foundation for Statistical Computing, 2024. Disponível em: <https://CRAN.R-project.org/package=car>. Acesso em: 9 set. 2025.

FREES, Edward W. **Regression Modeling with Actuarial and Financial Applications**. Cambridge: Cambridge University Press, 2010. ISBN 978-0-521-76011-9.

GAO, Lei; GUAN, Ling. **Interpretability of Machine Learning:** Recent Advances and Future Prospects. IEEE Multimedia, [s. l.], 2023. Disponível em: https://arxiv.org/abs/2305.00537. Acesso em: 29 jun. 2025.

GALVÃO, Alexandre M.; OLIVEIRA, Virgínia Izabel de; FLEURIET, Michel; e outros. **Gestão de riscos no mercado financeiro .** Rio de Janeiro: Saraiva Uni, 2018. E-book. pág.1. ISBN 9788547233037. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788547233037/. Acesso em: 26 jul. 2026.

GANIE, Mohd Waseem; WANI, Mohsin Altaf; WANI, Abid Hussain. **Machine learning approaches for crop yield prediction:** A systematic literature review. Remote Sensing Applications: Society and Environment, v. 26, 100284, 2026. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S2352938526002843>. Acesso em: 26 jul. 2026.

GEORGANOS, Stefanos; GRIPPA, Tais; VANHUYSSE, Sabine; LENNERT, Moritz; SHIMONI, Michal; WOLFF, Eleonore. **Very high resolution object-based land cover mapping with extreme gradient boosting**. IEEE Geoscience and Remote Sensing Letters, v.

15, n. 3, p. 1–5, 2018. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8304645>. Acesso em: 26 jul. 2026.

GUJARATI, Damodar N.; PORTER, Dawn C. **Econometria Básica.** 5. ed. Rio de Janeiro: Elsevier, 2011. Disponível em: <https://www.kufunda.net/publicdocs/ECONOMETRIA%20B%C3%81SICA%20-%20Gujarati.pdf>. Acesso em: 5 mar. 2026.

HANAFY, Mohamed; MING, Ruixing. **Machine Learning Approaches for Auto Insurance Big Data**. Risks, v. 9, n. 2, p. 42, 2021. Disponível em: https://www.mdpi.com/2227-9091/9/2/42. Acesso em: 21 ago. 2025.

HARIOM, Tatsat; SAHIL, Puri; BRAD, Lookabaugh. **Blueprints de aprendizado de máquina e ciência de dados para finanças: desenvolvendo desde estratégias de trades até robôs Advisors com Python** . Rio de Janeiro: Editora Alta Livros, 2024. E-book. pi ISBN 9788550821726. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788550821726/. Acesso em: 26 jul. 2026.

HASSIJA, Vikas; CHAMOLA, Vinay; MAHAPATRA, Atmesh; SINGAL, Abhinandan; GOEL, Divyansh; HUANG, Kaizhu; SCARDAPANE, Simone; SPINELLI, Indro; MAHMUD, Mufti; HUSSAIN, Amir. **Interpreting Black-Box Models:** A Review on Explainable Artificial Intelligence. Cognitive Computation, v. 16, p. 45–74, 2024. Disponível em: https://link.springer.com/article/10.1007/s12559-023-10179-8. Acesso em: 29 jun. 2025.

IBGE. **Estatísticas da produção agropecuária.** Rio de Janeiro: IBGE, 2024. Disponível em: <https://www.ibge.gov.br>. Acesso em: 1 jul. 2025.

IBGE. **Produção Agropecuária**. [*S. l.*], 20 abr. 2022. Disponível em: https://www.ibge.gov.br/explica/producao-agropecuaria/. Acesso em: 24 jul. 2024.

IBGE. Tabela 1737 – **IPCA:** série histórica com número-índice, variação mensal e variações acumuladas em 3 meses, em 6 meses, no ano e em 12 meses (a partir de dezembro/1979). SIDRA – Sistema IBGE de Recuperação Automática. Disponível em: <https://sidra.ibge.gov.br/Tabela/1737>. Acesso em: 7 mar. 2026.

INSTITUTO CLIMA E SOCIEDADE (ICS). **Brazil Must Protect its Agribusiness from Climate Impacts.** Rio de Janeiro: ICS, 2025. Disponível em: <https://climaesociedade.org/en/artigo/brazil-must-protect-its-agribusiness-from-climate-impacts>. Acesso em: 1 jul. 2025.

JAMES, Gareth; HARLAND, John H.; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert; TAYLOR, Jonathan. **An Introduction to Statistical Learning:** with Applications in R. 2. ed. New York: Springer, 2023. Disponível em: https://www.statlearning.com. Acesso em: 28 ago. 2025.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. **Forecasting:** Principles and Practice. 2. ed. Melbourne: OTexts, 2018.

KAMILARIS, Anastasios; PRENAFETA-BOLDÚ, Francesc. **Deep learning in agriculture: a survey.** Computers and Electronics in Agriculture, v. 147, p. 70–90, 2018. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0168169917308803>. Acesso em: 26 jul. 2026.

KUHN, Max; JOHNSON, Kjell. **Applied predictive modeling.** New York: Springer, 2013.

KHAN, Muhammad Yaseen; NIZAMI, Muhammad Suffian; SIDDIQUI, Muhammad Shoaib; WASI, Shaukat; *et al.* **Automated Prediction of Good Dictionary Examples (GDEX):** A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. Complexity, v. 2021, p. 1-15, set. 2021. DOI: 10.1155/2021/2553199. Disponível em: <https://www.researchgate.net/publication/354354484>. Acesso em: 6 mar. 2026.

KLUGMAN, Stuart A.; PANJER, Harry H.; WILLMOT, Gordon E. **Loss Models:** From Data to Decisions. 5. ed. Hoboken: Wiley, 2019.

KRATKA, Zuzana. **Mathematical-Statistical Models in Insurance**, European Scientific Journal, v. 11, p. 81-91, 2 mar. 2015. Disponível em https://eujournal.org/index.php/esj/article/view/5308/0?source=/index.php/esj/article/view/5308/0. Acesso em: 26 jul. 2024.

LIAKOS, J.; BUSATO, P.; MOSHOU, D.; PEARSON, S.; BOCHTIS, D. **Machine learning applications in agriculture: a review. Sensors,** v. 18, n. 7, 2074, 2018. Disponível em: <https://doi.org/10.3390/s18072074>. Acesso em: 26 jul. 2026.

LUCAS, Edimilson Costa; DI AGUSTINI, Carlos Alberto; SANTOS, Hayssa Guilherme; OLIVEIRA, José Carlos da Silva; RIBEIRO, Victor Hugo Almeida; NETO, Luiz Carlos. **Previsão de sinistros com machine learning:** evidências no setor de seguros automotivos no Brasil. Revista de Gestão e Secretariado, v. 16, n. 6, p. 1–20, 2025. Disponível em: <https://ojs.revistagesec.org.br/secretariado/article/view/4970>. Acesso em: 28 jun. 2025.

LUNDBERG, Scott M.; LEE, Su-In; *et al.* **SHAP Documentation**. [S. l.]: SHAP Developers, 2023. Disponível em: <https://shap.readthedocs.io/>. Acesso em: 17 nov. 2025.

MAISELI, Baraka Jacob. **Optimum design of chamfer masks using symmetric mean absolute percentage error**. EURASIP Journal on Image and Video Processing, v. 2019, n. 74, p. 1-15, 2019. Disponível em: <https://link.springer.com/content/pdf/10.1186/s13640-019-0475-y.pdf>. Acesso em: 17 nov. 2025.

MAHAJAN, Siddharth; AGARWAL, Rohan; GUPTA, Mihir. **Algorithmic Bias Under the EU AI Act:** Compliance Risk, Capital Strain, and Pricing Distortions in Life and Health Insurance Underwriting. *Risks*, v. 13, n. 9, p. 160, 2025. Disponível em: https://www.mdpi.com/2227-9091/13/9/160. Acesso em: 24 ago. 2025.

MARTINS, Norberto Montani. **Risco sistêmico, fragilidade financeira e crise: uma análise pós-keynesiana a partir da contribuição de Fernando Cardim de Carvalho.** Revista de Economia Contemporânea, Rio de Janeiro, v. 24, n. 2, p. 1-25, 2020. DOI: http://dx.doi.org/10.1590/198055272428. Disponível em: https://revistas.ufrj.br/index.php/rec/article/view/198055272428. Acesso em: 26 jul. 2026.

MASÍS, Serg. **Interpretable Machine Learning with Python**. 2. ed. Birmingham: Packt Publishing, 2023.

MOLNAR, Christoph. **Interpretable Machine Learning**: A Guide for Making Black Box Models Explainable. [*S. l.*: *s. n.*], 2024. Disponível em: https://christophm.github.io/interpretable-ml-book/. Acesso em: 28 jul. 2024.

MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. **Introduction to Linear Regression Analysis.** 5th ed. Hoboken, NJ: John Wiley & Sons, 2012. ISBN 978-0-470-54281-1.

MORETTIN, Pedro A. **Estatística básica** . 10. ed. Rio de Janeiro: Saraiva Uni, 2023. E-book. pi ISBN 9788571441484. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788571441484/. Acesso em: 26 jul. 2026.

MCKINNEY, Wes *et al.* **pandas:** Python Data Analysis Library. Versão 2.3.2. [S. l.]: pandas via NumFOCUS, Inc., 2025. Disponível em: <https://pandas.pydata.org/?hl=pt-BR>. Acesso em: 9 set. 2025.

NAVATHA, Bajjaboina. **Statistical modelling for insurance data.** International Journal of Mathematics and Statistics Invention, [S.l.], v. 10, n. 6, p. 1–4, nov./dez. 2022. Disponível em: https://www.ijmsi.org/Papers/Volume.10.Issue.6/A10060104.pdf. Acesso em: 12 jul. 2025.

NAGY, Cristina Mihaela; COTLEȚ, Dumitru. **Specific aspects of the technical reserves of insurance accounting.** Annals of the University of Petrosani – Economics, v. 11, n. 1, p. 171–178, 2011. Disponível em: https://www.upet.ro/annals/economics/pdf/2011/Nagy-Cotlet.pdf. Acesso em: 6 nov. 2025.

NETO, Alexandre A. **Matemática Financeira e suas aplicações** . 15. ed. Rio de Janeiro: Atlas, 2022. E-book. p.Capa. ISBN 9786559773244. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9786559773244/. Acesso em: 25 jul. 2026.

NOORUNNAHAR, Mst; CHOWDHURY, Arman Hossain; MILA, Farhana Arefeen. **A Tree-Based XGBoost Machine Learning Model to Forecast the Annual Rice Production in Bangladesh.** PLOS ONE, v. 18, n. 3, e0283452, 2023. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0283452. Acesso em: 24 ago. 2025.

NOLOGY. **Modelagem preditiva: o que é e como pode revolucionar seu negócio**. 24 out. 2024. Disponível em: https://nology.com.br/blog/o-que-e-modelagem-preditiva/. Acesso em: 11 nov. 2025.

NORMA FEDERAL - DISPÕE SOBRE O SEGURO RURAL E O FUNDO DE ESTABILIDADE DO SEGURO RURAL - FESR. Resolução CNSP nº 404, de 26 de março de 2021. A Superintendência de Seguros Privados - SUSEP, no uso da atribuição que lhe confere... [S. l.], 26 mar. 2021. Disponível em: https://www.normasbrasil.com.br/norma/resolucao-404-2021_411679.html. Acesso em: 18 jul. 2024.

NUMPY DEVELOPERS. **NumPy**. [S. l.]: NumPy Developers, 2023. Disponível em: <https://numpy.org/>. Acesso em: 17 nov. 2025.

OSSIFO, Momate Emate; NAKAMURA, Luiz Ricardo; PEDRO, César; MUCHICO, Joaquina da Márcia Jaime; FERREIRA, Daniel Furtado; SOUZA, João Cândido de; RIBEIRO, Alex de Oliveira. **Produtividade do milho com base em uma abordagem de regressão distribucional.** Pesquisa Agropecuária Brasileira, v. 59, e03690, 2024. Disponível em: https://apct.sede.embrapa.br/pab/article/view/27824. Acesso em: 20 ago. 2025.

PARGENT, Florian; PFISTERER, Florian; THOMAS, Janek; BISCHL, Bernd. **Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features.** Computational Statistics, v. 37, p. 2671–2692, 2022. DOI: <https://doi.org/10.1007/s00180-022-01207-6>. Disponível em: <https://www.researchgate.net/journal/Computational-Statistics-1613-9658/publication/359022074_Regularized_target_encoding_outperforms_traditional_methods_in_supervised_machine_learning_with_high_cardinality_features/links/6222cb5d97401151d2fd4ea6/Regularized-target-encoding-outperforms-traditional-methods-in-supervised-machine-learning-with-high-cardinality-features.pdf>. Acesso em: 7 mar. 2026.

PEDULLA, Nicola Rocco de Luna. **Interpretabilidade de modelos de aprendizagem de máquina no mercado de seguros.** 2022. Trabalho de Graduação (Bacharelado em Ciência da Computação) – Centro de Informática, Universidade Federal de Pernambuco, Recife, 2022. Disponível em: https://www.cin.ufpe.br/~tg/2022-1/tg_CC/tg_nrlp.pdf Acesso em: 4 fev. 2026.

PEREIRA, Ana Carolina; OLIVEIRA, Rodrigo Martins; SILVA, Thiago Fernandes; MENDES, Larissa Rodrigues; COSTA, João Pedro. **Enhancing crop insurance analysis with agricultural zoning data.** Revista de Economia e Sociologia Rural, v. 63, n. 1, 2025. Disponível em: <https://www.scielo.br/j/resr/a/R3xNJ5fPgmFdn4dDtJPCztc/>. Acesso em: 01 jul. 2025.

PÉREZ, Fernando Lucambio. **Modelos Lineares Generalizados.** Última atualização em: 22 maio 2022. Disponível em: <http://www.leg.ufpr.br/~lucambio/GLM/GLM.html>. Acesso em: 6 mar. 2026.

PYLE, Dorian. **Data preparation for data mining.** San Francisco: Morgan Kaufmann Publishers, 1999.

PYTHON SOFTWARE FOUNDATION. **Python:** linguagem de programação. Versão 3.13.7. Wilmington: Python Software Foundation, 2025. Disponível em: https://docs.python.org/pt-br/3.13/. Acesso em: 9 set. 2025

RIFALDI, D.; FAMUJI, T. S.; WIJAYA, S. A.; *et al.* **Machine Learning 5.0:** In-depth Analysis Trends in Classification. Scientific Journal of Computer Science, [S. l.], v. 1, n. 1, p. 1–10, 2025. Disponível em: https://journal.futuristech.co.id/index.php/sjcs/article/view/18/8. Acesso em: 29 jun. 2025.

RIPLEY, Brian; VENABLES, Bill. **MASS:** Support Functions and Datasets for Venables and Ripley's MASS. Versão 7.3-65. Vienna: R Foundation for Statistical Computing, 2025. Disponível em: <https://cran.r-project.org/package=MASS>. Acesso em: 9 set. 2025.

ROBINSON, David; HAYES, Alex; COUCH, Simon; POSIT SOFTWARE, PBC. **broom:** Convert Statistical Objects into Tidy Tibbles. Versão 1.0.9. Vienna: R Foundation for Statistical Computing, 2025. Disponível em: <https://CRAN.R-project.org/package=broom>. Acesso em: 9 set. 2025.

SEABOLD, Skipper; PERKTOLD, Josef. **Statsmodels:** Econometric and Statistical Modeling with Python. Version 0.14.0. [S. l.]: Python Software Foundation, 2010. Disponível em: https://www.statsmodels.org/. Acesso em: 8 mar. 2026.

SEGURO RURAL. [*S. l.*], 12 jul. 2022. Disponível em: https://www.gov.br/susep/pt-br/planos-e-produtos/seguros/seguro-rural. Acesso em: 17 jul. 2024.

SILVA, Leandro Augusto da; PERES, Sarajane Marques; BOSCARIOLI, Clodis. I**ntrodução à mineração de dados:** com aplicações em R. Rio de Janeiro: Elsevier, 2016.

SOUZA, Priscila; ASSUNÇÃO, João. **Gerenciamento de risco na agricultura brasileira:** instrumentos, políticas públicas e perspectivas. Brasília: Ministério da Agricultura, Pecuária e Abastecimento, 2020. Disponível em: https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/seguro-rural/observatorio-do-seguro-rural/estudos/estudos-2020/2020-priscila-souza-gerenciamento-de-risco-na-agricultura-brasileira-instrumentos-politicas-publicas-e-perspectivas.pdf. Acesso em: 6 nov. 2025.

SU, Xiaoshan; BAI, Manying. **Stochastic Gradient Boosting Frequency-Severity Model of Insurance Claims.** PLOS ONE, v. 15, n. 8, e0238000, 2020. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0238000. Acesso em: 24 ago. 2025.

SUPERINTENDÊNCIA DE SEGUROS PRIVADOS (SUSEP). Circular SUSEP nº 682, de 18 de dezembro de 2022. Dispõe sobre a codificação dos ramos de seguro e classificação das coberturas para fins de contabilização. Disponível em: <https://www2.susep.gov.br/safe/scripts/bnweb/bnmapi.exe?router=upload/26949>. Acesso em: 5 ago. 2026.

SCIKIT-LEARN DEVELOPERS. **scikit-learn:** Machine Learning in Python. Versão 1.7.1. [S. l.]: Scikit-learn Project, 2025. Disponível em: <https://scikit-learn.org/stable/install.html>. Acesso em: 9 set. 2025.

SCHMIDT, Anderson; WITTMANN, Milton Luiz; SEIBERT, Rosane Maria. **Estratégias de gestão na concessão de crédito: prevenção da inadimplência e recuperação de crédito em instituições financeiras.** 2026. Disponível em: https://doi.org/10.36557/2674-9432.2026v5n2p1213-1233. Acesso em: 26 jul. 2026.

SHAMS, Mostafa; GHOSH, Kaushik. **Modeling insurance claims using Bayesian nonparametric regression.** PLOS One, v. 21, n. 4, e0346734, 2026. Disponível em: <https://doi.org/10.1371/journal.pone.0346734>. Acesso em: 26 jul. 2026.

SHINDE, Aditya; RAUT, Purva. Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018). **Comparative Study of Regression Models and Deep Learning Models for Insurance Cost Prediction**, Vellore, India, v. 1, p. 1124-1133, 6 dez. 2018.

STAUDT, Yves; WAGNER, Joël. **Assessing the performance of random forests for modeling claim severity in collision car insurance.** Risks, Basel, v. 9, n. 3, p. 53, 2021. Disponível em: https://www.mdpi.com/2227-9091/9/3/53. Acesso em: 21 ago. 2025.

STEINER, Kimberly; MENG, Boyang. **Predictiveness vs. interpretability.** Society of Actuaries – Compact Newsletter, Schaumburg, IL, out. 2019. Disponível em: https://www.soa.org/globalassets/assets/library/newsletters/compact/2019/october/2019-compact-iss63-steiner-meng.pdf. Acesso em: 12 jul. 2025.

SWAIN, Dillip Kumar; MOHANTY, Uma Charan; SINHA, Palash; RAO, M. M. Nageswara; SINGH, Kripan Kumar. **Climate Risk Management in Agriculture:** Monthly and Seasonal

Forecast Application. Cham: Springer, 2025. Disponível em: <https://link.springer.com/book/10.1007/978-3-031-51862-1>. Acesso em: 01 jul. 2025.

TAERO, Élio José. **Estatística e análise do risco:** aplicações e ligações. 2016. Dissertação (Mestrado em Matemática, Estatística e Computação – Especialização em Estatística Computacional) – Universidade Aberta, Lisboa, 2016. Disponível em: https://repositorioaberto.uab.pt/bitstream/10400.2/5786/1/TMEMC_ElioTaero.pdf. Acesso em: 6 nov. 2025.

TANURE, Tarik Marques do Prado; ALVES, Everton Luiz; ASSAD, Eduardo Delgado; GHINI, Raquel; HAMADA, Emília. **Mudanças do clima e agropecuária:** impactos, mitigação e adaptação. Desafios e oportunidades. Estudos Avançados, São Paulo, v. 38, n. 112, p. 109–132, 2024. Disponível em: https://www.scielo.br/j/ea/a/jJP56TJd4ZCKvQ4YmPhXgCk/. Acesso em: 26 jun. 2025.

TANURE, Tarik Marques do Prado; DOMINGUES, Edson Paulo; MAGALHÃES, Aline Souza. **Regional impacts of climate change on agricultural productivity: evidence on large-scale and family farming in Brazil.** Revista de Economia e Sociologia Rural, v.62, n.1, 2024. DOI: 10.1590/1806-9479.2022.262515. Acesso em: 6 ago. 2026.

VANDERVORST, Félix; VERBEKE, Wouter; VERDONCK, Tim. **Data misrepresentation detection for insurance underwriting fraud prevention.** Decision Support Systems, v. 159, 113798, 2022. Disponível em: https://doi.org/10.1016/j.dss.2022.113798. Acesso em: 26 jul. 2026.

VERBEKE, R.; DEJAEGER, K.; MARTENS, D. **Predictive modeling in insurance: applications of machine learning techniques.** Expert Systems with Applications, v. 168, 114695, 2021. Disponível em: https://doi.org/10.1016/j.eswa.2021.114695. Acesso em: 26 jul. 2026.

VINUESA, R. *et al.* **Decoding complexity:** how machine learning is redefining scientific discovery. arXiv, [s. l.], 2025. Disponível em: https://arxiv.org/abs/2405.04161. Acesso em: 29 jun. 2025.

VOOO. **Um tutorial completo sobre modelagem baseada em árvores de decisão (códigos R e Python)**. 12 dez. 2016. Disponível em: https://www.vooo.pro/insights/um-tutorial-completo-sobre-a-modelagem-baseada-em-tree-arvore-do-zero-em-r-python/. . Acesso em: 11 nov. 2025.

WANG, Yuanchao; PAN, Zhichen; ZHENG, Jianhua; QIAN, Lei; LI, Mingtao. **A hybrid ensemble method for pulsar candidate classification**. Astrophysics and Space Science, v. 364, n. 8, 2019. Disponível em: https://www.researchgate.net/publication/335483097_A_hybrid_ensemble_method_for_pulsar_candidate_classification. Acesso em: 6 nov. 2025.

WICKHAM, Hadley *et al.* **ggplot2:** Create Elegant Data Visualisations Using the Grammar of Graphics. Versão 3.5.2. Vienna: R Foundation for Statistical Computing, 2025. Disponível em: <https://CRAN.R-project.org/package=ggplot2>. Acesso em: 9 set. 2025.

WOOLDRIDGE, Jeffrey M. **Introductory Econometrics:** A Modern Approach. 6. ed. Boston: Cengage Learning, 2016.

WHITE, Halbert. **A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.** Econometrica, v. 48, n. 4, p. 817-838, 1980.

XGBOOST DEVELOPERS. **XGBoost Documentation**. [S. l.]: XGBoost Developers, 2023. Disponível em: <https://xgboost.readthedocs.io/>. Acesso em: 17 nov. 2025.

YING, Xue. **An Overview of Overfitting and its Solutions.** Journal of Physics: Conference Series, 10 fev. 2019. Disponível em: https://www.researchgate.net/publication/331677125_An_Overview_of_Overfitting_and_its_Solutions. Acesso em: 28 jul. 2024.

ZHOU, Zhi-Hua. **Ensemble methods:** foundations and algorithms. Boca Raton: CRC Press, 2012.

ZHANG, Ruixun. **Toward interpretable machine learning: evaluating models of heterogeneous predictions.** Annals of Operations Research, v. 347, p. 867–887, abr. 2025.

ZHICHKIN, Kirill A.; NOSOV, Vladimir V.; ZHICHKINA, Lyudmila N. **Agricultural Insurance, Risk Management and Sustainable Development.** Agriculture (MDPI), Basel, v. 13, n. 7, p. 1317, jun. 2023. Disponível em: <https://www.researchgate.net/publication/372016206_Agricultural_Insurance_Risk_Management_and_Sustainable_Development>. Acesso em: 01 jul. 2025.

APÊNDICE CÓDIGO

Todos os códigos elaborados para execução deste trabalho podem ser encontrados no repositório GitHub: https://github.com/CrisBlue128/TCC