UNIVERSIDADE FEDERAL DO CEARÁ

CONTEXTUS
REVISTA CONTEMPORÂNEA DE ECONOMIA E GESTÃO
Contextus – Contemporary Journal of Economics and Management

ISSN 1678-2089
ISSNe 2178-9258

www.periodicos.ufc.br/contextus

Árvore de decisão aplicada na classificação de ocorrência de sinistro cibernético em empresas do setor bancário

Decision tree applied in classifying the occurrence of cyber claims in banking sector companies

Árbol de decisión aplicado en la clasificación de la ocurrencia de siniestros cibernéticos en empresas del sector bancario

https://doi.org/10.19094/contextus.2023.e83423

Alana Katielli Nogueira Azevedo
https://orcid.org/0000-0002-3700-4916
Professora na Universidade Federal do Ceará (UFC)
Doutoranda em Matemática Aplicada à Economia e à Gestão na Universidade de Lisboa (ULISBOA)
Mestre em Economia pela Universidade Federal do Ceará (UFC)
alanakna@gmail.com

RESUMO
O estudo teve como objetivo a previsão de sinistros cibernéticos em empresas do setor bancário através do uso de árvore de decisão. Para tanto, foram extraídos 683 casos de perdas cibernéticas de um banco de dados de risco operacional. As variáveis independentes consideradas na modelagem foram a região de domicílio, o porte da empresa e, como principal variável explicativa, o faturamento. A classificação apresentou 89% de acertos globais. A modelagem em questão garante uma boa qualidade de classificação e melhor ajuste quando comparada a modelagem tradicional GLM. Os resultados desse trabalho são úteis e podem atuar de forma inovadora como ferramenta de apoio à tomada de decisão das seguradoras, visando respostas úteis ao gerenciamento de riscos cibernéticos.
Palavras-chave: gerenciamento de risco; risco cibernético; árvore de decisão; GLM; setor bancário.

ABSTRACT
The study aimed to predict cyber claims in companies in the banking sector using a decision tree. To this end, 683 cases of cyber losses were extracted from an operational risk database. The independent variables considered in the modeling were the region of domicile, the size of the company and, as main explanatory variable, revenue. The classification reached 89% of global hits. The modeling in question guarantees a good classification quality and better fit when compared to traditional GLM modeling. The results of this work are useful and can act in an innovative way as a tool to support the decision making of insurers, aiming at useful responses to the management of cyber risks.
Keywords: risk management; cyber risk; decision tree; GLM; banking sector.

RESUMEN
El estudio tuvo como objetivo predecir ciber siniestros en empresas del sector bancario utilizando un árbol de decisión. Para ello, se extrajeron de una base de datos de riesgo operacional 683 casos de ciberpérdidas. Las variables independientes consideradas en la modelación fueron la región de domicilio, el tamaño de la empresa y, como principal variable explicativa, los ingresos. La clasificación alcanzó 89% de los hits globales. El modelado en cuestión garantiza una buena calidad de clasificación y un mejor ajuste en comparación con el modelado GLM tradicional. Los resultados son útiles y pueden actuar de forma innovadora como una herramienta de apoyo a la toma de decisiones de las aseguradoras, buscando respuestas útiles a la gestión de los riesgos cibernéticos.
Palabras clave: gestión de riesgos; ciberriesgo; árbol de decisiones; GLM; sector bancario.

Informações sobre o Artigo
Submetido em 03/02/2023
Versão final em 04/04/2023
Aceito em 10/04/2023
Publicado online em 17/10/2023

Comitê Científico Interinstitucional
Editor-Chefe: Diego de Queiroz Machado
Editora Adjunta: Alane Siqueira Rocha
Avaliado pelo sistema *double blind review* (SEER/OJS – versão 3)

Como citar este artigo:
Azevedo, A. K. N. (2023). Árvore de decisão aplicada na classificação de ocorrência de sinistro cibernético em empresas do setor bancário. *Contextus – Revista Contemporânea de Economia e Gestão*, 21(esp.1), e83423. https://doi.org/10.19094/contextus.2023.e83423

1 INTRODUÇÃO

A era digital veio para transformar definitivamente as estruturas corporativas e desenvolver tecnologias de informação complexas. Por outro lado, tal evolução acarreta vulnerabilidades quando se trata de ameaças cibernéticas. Ataques cibernéticos são cada vez mais constantes e podem gerar prejuízos de ordem financeira em grau muito elevado.

Segundo Allianz (2022), riscos cibernéticos são a principal preocupação para empresas ao redor do mundo. O Brasil, no ano de 2022, ficou em segundo lugar no ranking de países que mais sofreram ataques cibernéticos na América Latina, ficando atrás apenas do México e apresentando um aumento de 94% em relação ao ano de 2021 (Fortinet, 2022). Ataques hackers, violação de dados e falhas nos sistemas estão entre as principais ameaças. Melhorar a compreensão sobre esse tipo de risco é o atual desafio para gestores.

Em se tratando de serviços financeiros, estes sempre foram muito visados quando se fala de fraudes. Com o advento de novos meios de pagamento, formas de relação cliente/empresa cada vez mais digitalizadas, possibilidades oferecidas pelo open banking, tal setor se tornou ainda mais atrativo para criminosos cibernéticos. Desta forma, a gestão de vulnerabilidade se torna essencial. Identificar, notificar, analisar e corrigir as vulnerabilidades de segurança cibernética fazem parte desse processo (Ecotrust, 2023).

Aliado à gestão de vulnerabilidade, o seguro contra riscos cibernéticos se apresenta como opção de mecanismo de transferência de riscos. Apesar de estar ganhando atenção no mercado segurador, esse tipo de seguro ainda provoca questionamentos em relação a viabilidade financeira tanto para segurado como para segurador. Por se tratar de uma categoria de risco em ascensão, com disponibilização limitada de dados históricos, a precificação pode gerar valores por vezes imprecisos (Carfora, 2019).

O processo de modelagem de dados é parte fundamental para uma correta precificação no âmbito de seguros. Em contraste com o fato de que muitos estudos, dentre eles o de Carfora (2019) e Karam (2014), foram conduzidos para caracterizar e modelar o risco cibernético através de abordagens tradicionais como a teoria do risco coletivo, distribuição das perdas agregadas (LDA) e modelos lineares generalizados (GLM), este trabalho visa uma modelagem quantitativa com uso de aprendizagem de máquina, mais especificamente, árvores de decisão, metodologia que desenvolve algoritmos cada vez mais eficazes e eficientes, oferecendo a possibilidade de aumento da compreensão sobre o assunto em análise (Faceli, 2011).

Este trabalho propõe uma análise da frequência (número de ocorrências) de risco cibernético, utilizando todas as informações disponíveis de empresas do setor bancário e introduzindo uma estrutura de árvore de decisão capaz de identificar se uma determinada empresa está sujeita a sinistros cibernéticos. Os pontos fortes e diferenciais desse trabalho são: (i) No âmbito da metodologia de árvores de decisão, o uso de classes de risco para comparar com o tradicional GLM, identificando variáveis significativas de classificação de risco e (ii) O uso de dados reais de uma coleção mundial de perdas operacionais relatadas publicamente.

O manuscrito está organizado da seguinte forma. Na próxima seção, foi realizada uma revisão da literatura sobre os dois eixos temáticos, risco cibernético e árvore de decisão. A seção 3 é dedicada à apresentação da base de dados, como também descreve a metodologia de árvore de decisão aplicada para a previsão. Na seção 4 são expostos e discutidos os principais resultados. A última seção encerra o estudo com algumas considerações finais.

2 REFERENCIAL TEÓRICO

2.1 Riscos cibernéticos

A dependência de empresas dos diversos setores da economia em relação a tecnologias, principalmente as que gerem e armazenam informações por vezes valiosas, evidencia a necessidade de um correto gerenciamento de riscos. Dentre estes riscos, o risco cibernético vem ganhando destaque e, particularmente, as instituições financeiras estão cada vez mais conscientes das ameaças que esse tipo de risco pode trazer.

Segundo Dal Moro (2020), o risco cibernético geralmente se refere a qualquer risco de perda financeira, interrupção ou dano à reputação de uma organização, resultante da falha de seus sistemas de tecnologia da informação. Uma classificação adequada e a escolha de uma metodologia de gerenciamento pertinente a essa classe de risco são essenciais para que o processo de mitigação seja eficiente. Estudos vêm sendo desenvolvidos das mais variadas formas no intuito de se entender melhor as características dos riscos cibernéticos.

Peng et al. (2018) desenvolveram a primeira abordagem estatística, centrada em um modelo Cópula-GARCH que usa cópulas para modelar a dependência multivariada exibida por dados de ataques cibernéticos do mundo real. Tal metodologia é caracterizada por sua flexibilidade em poder acomodar diferentes estruturas de dependência entre diferentes pares de variáveis e capacidade de estimar um grande número de parâmetros. Os resultados mostram que a dependência multivariada entre ataques cibernéticos tem um efeito significativo na perda total. Os autores evidenciaram que ignorar a devida dependência multivariada causa uma subestimação severa dos riscos de segurança cibernética.

Xu e Hua (2019) produziram uma abordagem robusta e sistemática para modelar e precificar os riscos cibernéticos, estudando os riscos por meio de modelos epidêmicos, juntamente com as funções de perda e estratégias de preços. Os autores usaram processos estocásticos (Markov e não-Markov) para descrever a dinâmica de uma epidemia espalhada ao longo do tempo. Foi implementada uma abordagem de simulação para calcular o prêmio pelo risco de segurança cibernética para uso prático. Os efeitos de diferentes distribuições de infecção e dependência entre os processos de infecção nas perdas também foram estudados.

Subroto e Apriyana (2019) apresentaram um modelo algorítmico que utiliza análise de big data de mídia social e aprendizagem de máquina para prever riscos cibernéticos. Os dados para o estudo consistiram em 83.015 instâncias do banco de dados de vulnerabilidades e exposições comuns e 25.599 casos de riscos cibernéticos do Twitter. Considerando rede neural artificial e analisando as vulnerabilidades de software à ameaças, a experimentação resultou em uma taxa de precisão para a previsão de sinistros de 96,73%.

Carfora et al. (2019) apontaram as peculiaridades dos contratos de seguro cibernético em relação aos seguros não vida clássicos, tanto na perspectiva da seguradora quanto na perspectiva do segurado. As distribuições mais adequadas para representar a frequência (binomial negativa) e a severidade (log-normal) dos sinistros cibernéticos relatados são examinadas e a medida de Value at Risk foi estimada.

Muito se tem estudado a respeito de riscos cibernéticos, mas a falta de dados limita a capacidade do setor de seguros de propor cobertura para este tipo de risco. Marotta et al. (2017) explicam que as organizações têm medo de divulgar muitas informações sobre seus sistemas internos para evitar a diminuição da reputação, bem como evitar o vazamento de conhecimento sobre os pontos fracos do sistema. Para Eling e Schnell (2016), as dificuldades para segurar o risco cibernético são imensas, especialmente devido à falta de dados e abordagens de modelagem, riscos de acumulação incalculáveis e o risco de mudança que está atrelado à recursos ou orçamento inadequados, resistência à mudança de cultura organizacional, falta de apoio da gestão para a mudança e falta de compromisso com a mudança.

2.2 Árvores de decisão

Para problemas de regressão não paramétrica, árvores de decisão são uma ferramenta extremamente popular para obter previsões de alta qualidade (Linero, 2018). O uso de tal metodologia é vasto e variado, podendo ser utilizado em diversas áreas do conhecimento. Por exemplo, Hamoud et al. (2018) apresentaram um modelo baseado em algoritmos de árvore de decisão para analisar as informações coletivas de alunos de ensino superior como também classificar os dados coletados para prever e categorizar o desempenho do aluno.

Yuvaraj et al. (2021) criaram um modelo de árvore de decisão para classificação e identificação de textos com características de cyberbullying, uma epidemia entre os jovens. Já Bonini (2016) fez uso de árvore de decisão para extrair informações provenientes de uma base de dados de amostra de tumores de mama com intuito de realizar a classificação dos mesmos em benignos ou malignos.

Já para o setor financeiro, pesquisas vêm sendo realizadas com diversos propósitos. O modelo de previsão baseado em árvore de decisão proposto por Podhorská et al. (2020) auxilia na classificação adequada de empresas que possam ter dificuldades financeiras, chegando a falência, sob as condições dos mercados emergentes. Sousa et al. (2021), por sua vez, aplicaram três tipos de árvores de decisão para prever pagamentos de faturas. O primeiro modelo objetivava identificar faturas com pagamento no prazo ou atrasado. O segundo identificava, entre as faturas atrasadas, o pagamento no mês do vencimento ou posterior. Já o terceiro modelo previa, entre as faturas atrasadas, quantos dias de atraso teriam além do mês do vencimento. A precisão média obtida para os três modelos foi de 81,85%, 85,63% e 73,98%, respectivamente.

O artigo de RL e Mishra (2022) abordou a aplicação de algoritmos de árvore de decisão para prever o desempenho de empresas de manufatura em uma economia emergente. O estudo usa dados de 25 variáveis financeiras para uma amostra de 1.923 empresas manufatureiras indianas no período entre 2011 e 2018. Os resultados mostraram que a margem de lucro líquido e a taxa de rotatividade total de ativos são os fatores mais críticos que determinam o desempenho da empresa em um mercado indiano. Essas descobertas podem auxiliar os gerentes em seu processo de tomada de decisão e também ter implicações vitais para os investidores na avaliação do desempenho da empresa.

Saha et al. (2023) investigaram a questão de prever o desempenho financeiro de empresas de manufatura registradas em países em desenvolvimento usando métodos de aprendizado de máquina. O modelo teve média variando de 0,922 a 0,934 para previsão de vendas. Uma variável independente importante e significativa para prever as vendas em todas as categorias e algoritmos foram as despesas reais com matéria-prima, explicando aproximadamente 83% a 88% das somas totais de quadrados em todas as validações. A variável dependente lucros foi mais difícil de prever em relação às vendas. Segundo os autores, os resultados de uma abordagem de aprendizado de máquina podem aprimorar a compreensão dos mecanismos que traduzem as vendas em lucros.

Sembiring et al. (2021) fizeram uso de árvores de decisão com o objetivo de classificar clientes em prováveis solventes ou insolventes em relação à tomada de crédito bancário. Para os autores, uma boa seleção de clientes é indispensável para que estes possam pagar suas dívidas no prazo correto.

O fato de não assumir nenhuma distribuição particular para os dados, de poder considerar atributos tanto categóricos (qualitativos) como numéricos (quantitativos), de poder construir modelos para qualquer função desde que o número de exemplos de treinamento seja suficiente e de ter elevado grau de compreensão torna a utilização de árvores de decisão vantajosa (Lemos et al., 2005).

3 PROCEDIMENTOS METODOLÓGICOS

3.1 Base de dados

Em termos de mercado de seguros e setor financeiro, o risco cibernético é categorizado como risco operacional. Karam (2014) define essa categoria como o risco de perda decorrente de processos internos, pessoas e sistemas inadequados ou falhos ou de eventos externos. A enumeração de CRO (2016) ajudou a nortear a identificação dos riscos cibernéticos para esse trabalho, a saber:

- Quaisquer riscos decorrentes do uso de dados eletrônicos e sua transmissão, incluindo ferramentas tecnológicas como internet e redes de telecomunicações.
- Danos físicos que podem ser causados por ataques cibernéticos.
- Fraudes cometidas por uso indevido de dados.
- Qualquer responsabilidade decorrente do uso, armazenamento e transferência de dados.
- A disponibilidade, integridade e confidencialidade das informações eletrônicas (seja relacionadas a indivíduos, empresas ou governos).

Para a análise de risco cibernético, contou-se com o SAS OpRisk Global Data, que é a maior coleção do mundo de perdas operacionais divulgadas publicamente, organizado e fornecido pela empresa Statistical Analysis System. A base de dados considerada fornece informações sobre 26.762 ocorrências de perda operacional no período entre janeiro de 2004 e janeiro de 2021. Para cada ocorrência, a base de dados informa, além do valor da perda, a descrição do evento, as linhas de negócios e setores da indústria, a categoria do risco, país do incidente (que pode ser todo o mundo) e outras informações sobre as empresas envolvidas. Todas as perdas, expressas em US$, são apresentadas em valor presente, referente a janeiro de 2021, para a devida comparação.

O presente estudo considerou duas subcategorias para risco cibernético: (1) Ações de pessoas e (2) Falhas técnicas de sistemas. Considerando as informações do banco de dados do SAS com registros completos, foi identificado um total de 683 incidentes de risco cibernético em um conjunto de 2718 empresas do setor bancário. Os atributos escolhidos foram a região de domicílio, o porte da empresa de acordo com o número de empregados e faturamento em US$, determinantes importantes dos termos e preços de apólices de seguro cibernético, como destaca Biener et al. (2015).

3.2 Predição com o uso de árvores de decisão

Árvores de decisão, através da aprendizagem de máquina, oferecem um gama de algoritmos que dão suporte a modelos preditivos tanto para classificação como para regressão. Segundo Breiman et al. (2017), a ideia é representar dados como uma árvore onde cada nó interno representa um teste em um atributo, cada ramo representa um resultado do teste e cada nó folha apresenta um rótulo de classe.

O método de árvore de decisão é caracterizado por ser não paramétrico e supervisionado. Pela definição de Cunningham et al. (2008), o aprendizado supervisionado envolve aprender um mapeamento entre um conjunto de variáveis de entrada e uma variável de saída, aplicando esse mapeamento para prever as saídas para dados não vistos.

Existem dois tipos de árvores de decisão que se definem de acordo com a variável de saída. Quando esta for uma variável categórica, trata-se de uma árvore de decisão para classificação. Quando a variável de saída for contínua, refere-se a uma árvore de decisão para regressão.

A Figura 1 apresenta a estrutura básica de uma árvore de decisão. Cada árvore tem um nó raiz, por onde as entradas são passadas. Este nó raiz é dividido em conjuntos de nós de decisão onde os resultados e observações são baseados condicionalmente. Se um nó não se divide em mais nós, ele é chamado de nó folha ou nó terminal. Uma sub-seção de uma árvore de decisão é chamada de ramo ou sub-árvore.

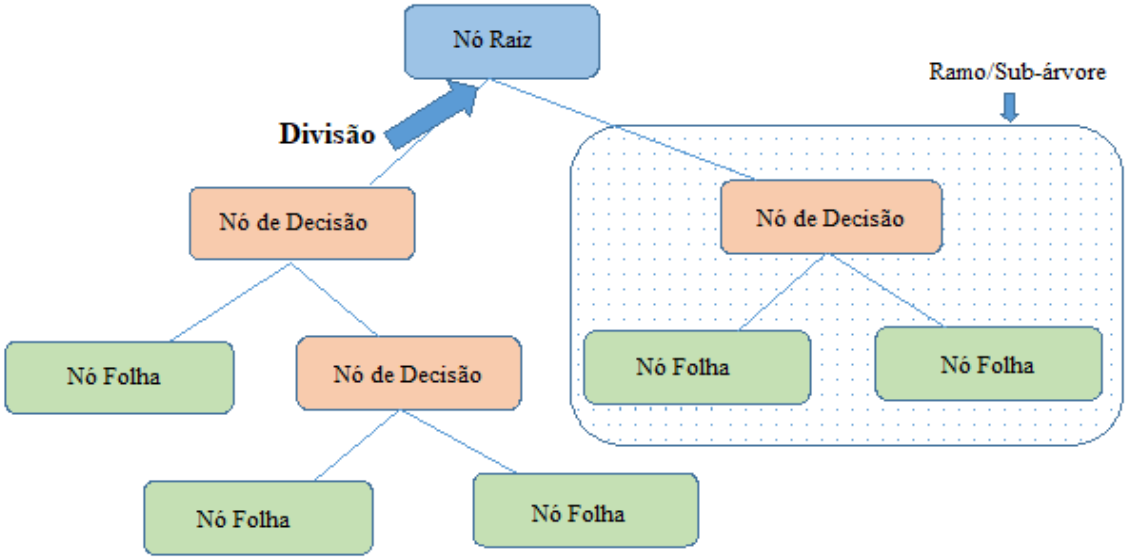

Figura 1. Exemplo de árvore de decisão.
Fonte: adaptado de Vidhya (2021).

O algoritmo de um modelo de árvore de decisão irá depender da variável de saída. No presente estudo, tal variável é categórica, o que determina o algoritmo a ser usado para classificação. Outra definição importante é como os dados sofrerão divisão. As regras de divisão influenciam a otimização e a performance do modelo.

A regra de divisão escolhida para ser aplicada na árvore de decisão aqui modelada foi a impureza de Gini. Segundo Ruiz-Maya (1978), o objetivo da impureza de Gini é medir o grau de importância de cada variável explicativa. No âmbito de árvores de decisão, tal regra é usada para medir a probabilidade de um exemplo escolhido aleatoriamente ser classificado erroneamente por um determinado nó. Quando todos os elementos são corretamente divididos em diferentes classes, a divisão é considerada pura.

A Equação 1 representa matematicamente a medida de impureza de Gini, onde pi é a probabilidade de um determinado elemento pertencer a uma classe específica.

$$Gini = 1 - \sum_{i=1}^{n}(p_i)^2 \quad (1)$$

A pontuação de impureza de Gini concentra valores entre 0 e 1. Quando esta for igual a 0 a divisão é chamada de pura, desta forma, todos os elementos pertencem a uma determinada classe. Em caso de valor igual a 1 os elementos estão segregados aleatoriamente em diversas classes.

3.3 Medidas de desempenho

Uma das formas de representação para verificar o desempenho do modelo de árvore de decisão é a matriz de confusão, já que se trata de um problema de duas classes. Classifica-se uma classe como sendo positiva (+) e outra como negativa (-). O modelo matricial pode ser visto na Tabela 1, onde:

- VP corresponde ao número de empresas que sofreram sinistros cibernéticos e foram classificadas como tal.
- VN corresponde ao número de empresas que não sofreram sinistros cibernéticos e foram classificadas como tal.
- FP corresponde ao número de empresas que sofreram sinistros cibernéticos e foram classificadas como empresas sem sinistros.
- FN corresponde ao número de empresas que não sofreram sinistros cibernéticos e foram classificadas como empresas que sofreram sinistros.

Tabela 1
Matriz de confusão para problemas de duas classes

	Valores preditos	
Valores reais	+	-
+	VP	FN
-	FP	VN

Fonte: Elaboração própria.

A partir da matriz de confusão, outras medidas podem ser calculadas para avaliar a eficácia do modelo de árvore de decisão. Neste trabalho serão calculadas a taxa de erro total, acurácia total, sensibilidade e especificidade.

A taxa de erro total, Equação 2, é representada pela soma da diagonal principal da matriz de confusão, dividida pela soma de todos os elementos da matriz. Acurácia é a medida que traduz a precisão de um teste (Equação 3).

$$err = \frac{FP+FN}{VP+FP+VN+FN} \quad (2)$$

$$ac = \frac{VP+VN}{VP+FP+VN+FN} \quad (3)$$

De acordo com Martinez et al. (2003), a sensibilidade (Equação 4) é a probabilidade de o teste sob análise fornecer resultado positivo, ou seja, traduz a capacidade do teste de identificar uma empresa que está sofrendo sinistro cibernético. Ainda segundo o autor, a especificidade (Equação 5) é a probabilidade de o teste fornecer resultado negativo, traduzindo a capacidade do teste de identificar uma empresa que não sofre sinistro cibernético.

$$sens = \frac{VP}{VP+FN} \quad (4)$$

$$esp = \frac{VN}{VN+FP} \quad (5)$$

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

4.1 Análise exploratória de dados

A Tabela 2 fornece um resumo da amostra de risco cibernético. Em relação à região geográfica, o Painel B mostra que as empresas da América do Norte apresentam quase a metade dos incidentes (48,9%). A Europa vem em segundo lugar com 25,3%. Apesar de uma maior quantidade de sinistros, em relação ao valor médio das perdas, a América do Norte apresenta uma das menores. Para Biener et al. (2015), as empresas norte-americanas são mais capazes e dispostas a investir em medidas de mitigação de risco para perdas extremas.

Separando as empresas por tamanho com base em quantis, o Painel C da Tabela 2 mostra uma semelhança no número de sinistros em cada uma das categorias, apesar das diferenças em relação ao faturamento médio de cada grupo. Vê-se também que a chance de se ter um sinistro cibernético é maior nas empresas de grande porte, com probabilidade de 50,6%, percentual bem mais elevado quando comparado com as empresas de pequeno (14,2%) e médio porte (37%).

Ainda analisando o Painel C, verifica-se que o valor médio dos sinistros é bem semelhante para todos os tipos de porte de empresas. Para ClearSale (2022), pequenas e médias empresas podem ser consideradas alvos fáceis para criminosos que buscam mais agilidade nos golpes devido à inexperiência. Fato que pode gerar prejuízos elevados.

Tabela 2
Caracterização das empresas do setor bancário

	Nº de Empresas	Faturamento médio das empresa (em milhões US$)	Nº Empresas vítimas de sinistro cibernético	Valor médio dos sinistros (em milhões US$)
Painel A: Amostra total				
Total	2718	14 974,64	683	14,16
Painel B: Região de domicílio				
Asia	592	6 012,92	106	19,05
Europa	682	27 659,09	173	18,93
América do Norte	1.230	13 951,55	334	11,77
Outra	214	5 222,02	70	6,32
Painel C: Porte da empresa de acordo com o número de empregados				
Pequeno	1.679	1 818,70	238	15,86
Médio	592	12 024,07	219	12,15
Grande	447	68 298,01	226	14,31

Fonte: Elaboração própria.
Legenda: a classificação por porte é baseada nos quantis inferior, médio e superior de 33% do número de funcionários; Pequeno (≤ 7.100 funcionários); Médio (entre 7.170 e 56.137 funcionários); Grande (≥ 56.218 funcionários).

A Tabela 3 mostra que, em 89,4% dos casos, o comportamento humano é a principal fonte de incidentes de risco cibernético. Roubo de informações, danos causados por hackers e perda de dados de clientes são alguns exemplos. Em relação ao valor médio dos sinistros a situação é bastante diferente para as duas categorias. Os sinistros causados por falhas técnicas de sistemas geram 34,41 milhões de US$ a mais em prejuízos.

Tabela 3
Classificação dos sinistros cibernéticos

Ação geradora do sinistro cibernético	Nº de Empresas afetadas	Faturamento médio da empresa (em milhões US$)	Valor médio dos sinistros (em milhões US$)
Ações de pessoas	611	34 952,63	10,53
Falhas técnicas de sistemas	72	27 211,91	44,94

Fonte: Elaboração própria.

Aliada ao objetivo central desse estudo que é prever a ocorrência de sinistros em empresas do setor bancário e para se compreender melhor a categoria de risco cibernético, uma análise adicional da frequência desses sinistros foi realizada. Ajustar uma distribuição de probabilidade aos dados mensais do número de sinistros ajuda na análise do comportamento probabilístico desse risco. A Figura 2 mostra graficamente, através de um box-plot, a distribuição de ocorrência de sinistro cibernético no período considerado. A localização da maioria dos dados na parte inferior do gráfico é indicativa de assimetria, o que determina que os dados não podem ser normalmente distribuídos. Tal representação auxilia também na identificação de outliers, valores de dados que estão distantes dos outros e que podem afetar os resultados. Sua representação é realizada através de asteriscos, o que não foi o caso no presente estudo.

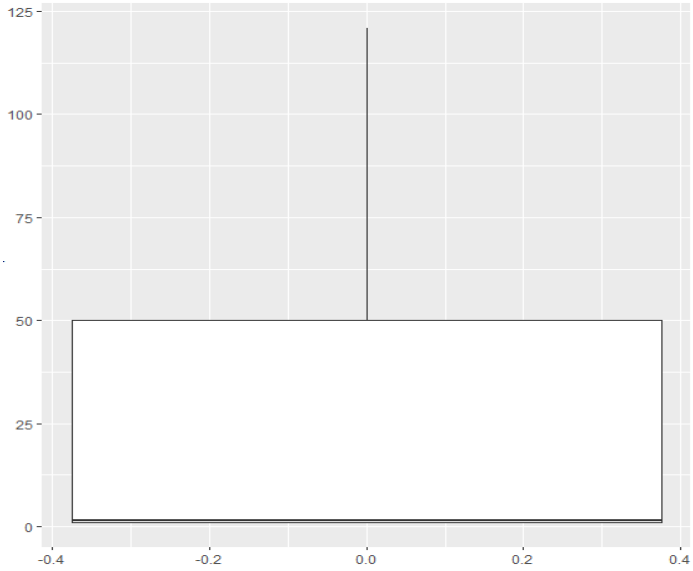

Figura 2. Box-plot representativo da frequência de sinistros.
Fonte: Elaboração própria.

Com o intuito de encontrar a melhor distribuição para representar a frequência de sinistros, considerou-se o critério de informação Akaike (AIC), um método que permite comparar modelos com diferentes famílias de distribuições e que não necessita de maiores inferências sobre o modelo para corroborar seu resultado (Burnham & Anderson, 2004). O melhor modelo é aquele com o menor valor de AIC. Para os dados aqui analisados, a distribuição logarítmica (LG) forneceu o melhor ajuste com um AIC de 182,52, além de ter parâmetro estimado estatisticamente significativo.

A Figura 3 mostra o histograma do ajuste e o worm plot, que fornece um diagnóstico sobre os resíduos. A média e variância dos resíduos foram -0,34 e 1,71, respectivamente, o que mostra um bom ajuste para a distribuição LG já que tais valores não se distanciam tanto em relação aos valores de uma distribuição normal padrão.

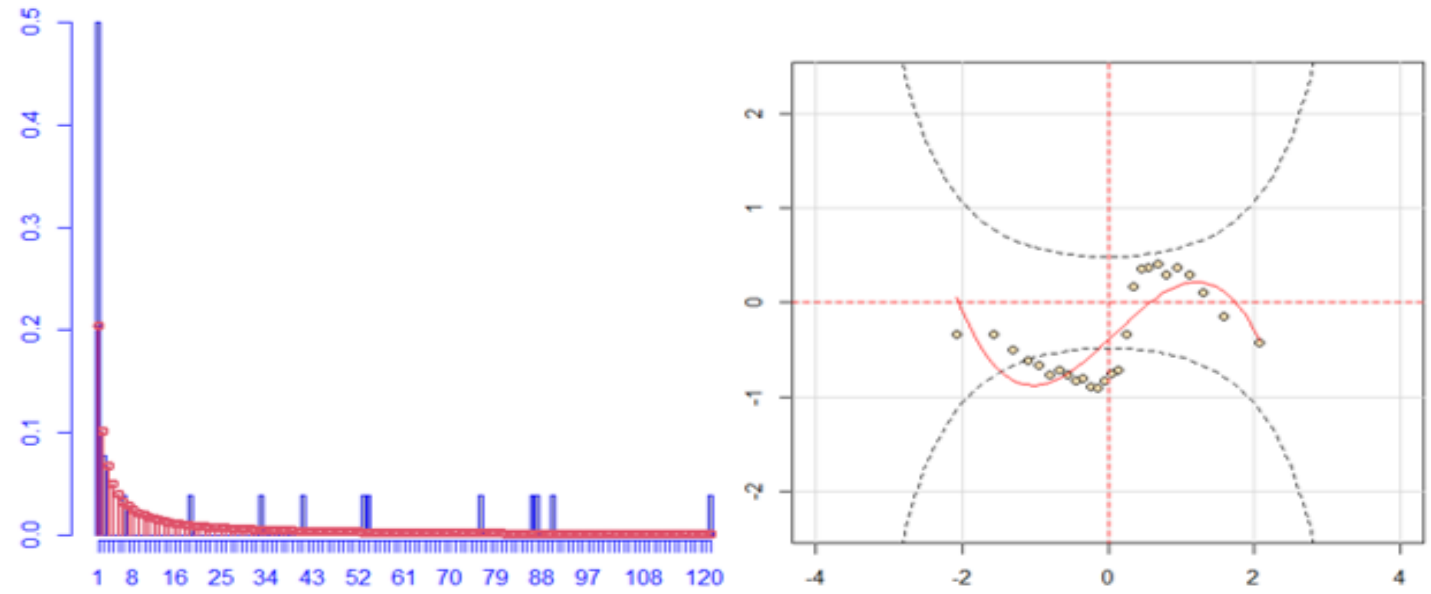

Figura 3. Histograma e worm plot do ajuste de distribuição LG para a frequência de sinistros.
Fonte: Elaboração própria.

O parâmetro da Tabela 4 foi estimado usando o método de máxima verossimilhança (ML), ver Portugal (1995). Esta informação é essencial para o cálculo do número esperado de sinistros, representado pelo primeiro momento da distribuição LG, definido pela Equação 6.

$$E[LG] = \frac{\beta}{(\beta-1)\ln(1-\beta)} \quad (6)$$

Tabela 4
Ajuste de distribuição LG para a frequência de sinistros

Parâmetro	Valor estimado	Erro padrão	Valor t	Pr(>|t|)
β	0,9922226	0,486406	9,96847	< 2,22e-16 ***

Fonte: Elaboração própria.
Legenda: códigos de significância = 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1.

Ao substituir o valor estimado do parâmetro β na Equação 1, totaliza-se o valor de 26,3 referente ao número esperado de sinistros de risco cibernético, considerando os dados mensais de ocorrência.

Toda essa informação sobre a frequência pode ser norteadora no processo de aceitação de risco pelas seguradoras, como também para maior conscientização acerca do risco cibernético a que empresas estão expostas.

4.2 Modelagem da árvore de decisão

A árvore definiu como principal variável explicativa o faturamento das empresas do setor bancário. Em seguida, a variável porte da empresa foi selecionada e a terceira variável foi a região de domicílio. Segundo Quinlan (1993), a variável mais importante é aquela com menor entropia e apresenta o maior ganho de informação. Tal importância é evidenciada na Figura 4.

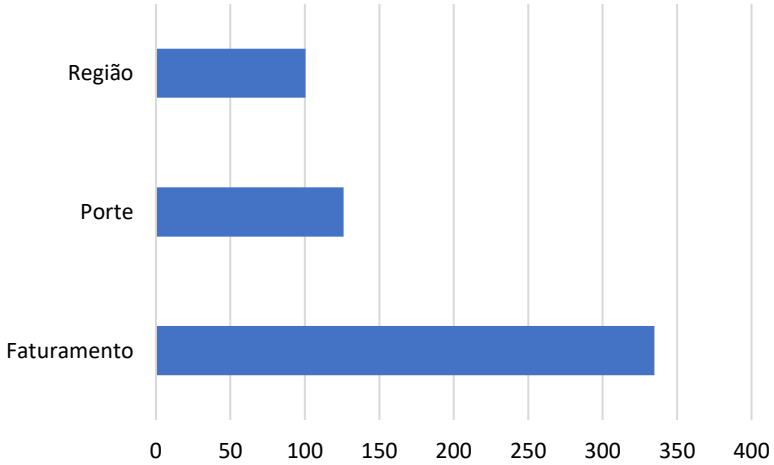

Figura 4. Importância das variáveis independentes.
Fonte: Elaboração própria.

O processo de classificação realizado na intenção de prever a ocorrência de sinistros cibernéticos, gerou os resultados apresentados na Tabela 5, como também a árvore de decisão exposta através da Figura 5.

Tabela 5
Matriz de confusão para classificação de sinistro cibernético

	Valores preditos	
Valores reais	+	-
+	422	261
-	97	1938

Fonte: Elaboração própria.

A proporção de concordância total (Acurácia) para a árvore de decisão foi de 89%, isto é, empresas que sofreram sinistros cibernéticos, como também empresas que não sofreram, foram classificadas corretamente em 89% dos casos.

Tabela 6
Medidas de desempenho do modelo de árvore de decisão

Medida de desempenho	**Árvore de decisão**	**GLM**
Acurácia	89%	78%
Erro	11%	22%
Sensibilidade	62%	28%
Especificidade	95%	95%

Fonte: Elaboração própria.

De acordo com os resultados apresentados na Tabela 6, observa-se que a árvore de decisão apresenta valor de sensibilidade de 62% e especificidade de 95%, mostrando que o modelo se mostrou mais eficiente em classificar a classe negativa do que a positiva. Apesar desta diferença percentual sempre há vantagens no uso da técnica de árvores de decisão, no sentido de que ela apresenta resultados de fácil compreensão, detalhando quais das informações sobre as empresas analisadas foram mais relevantes na classificação (Lemos et al., 2005).

Como pode ser visto, quando se compara o modelo de árvore de decisão com a abordagem GLM, houve queda de acurácia resultando em um erro com o dobro do valor. A taxa de sensibilidade diminuiu para 28%, enquanto a especificidade atingiu mesmo patamar de 95%. Tais valores indicam que o GLM implementado é muito bom na classificação de padrões pertencentes à classe negativa, mas perdeu eficiência para dados da classe positiva.

A diferença de performance entre as duas metodologias pode ser explicada considerando também o fato de que nos modelos GLM é necessária a escolha de uma distribuição particular da família exponencial para a variável resposta (Pekár & Brabec, 2017). Pelo ajuste realizado na seção 3.1 fica clara a impossibilidade de uso de GLM para análise dos dados de frequência de sinistros cibernéticos aqui apresentados ao se destacar a distribuição logarítmica como melhor aderência.

O fato é que o uso e estudo de árvores de decisão podem ser oportunos para auxiliar na resolução do problema apresentado.

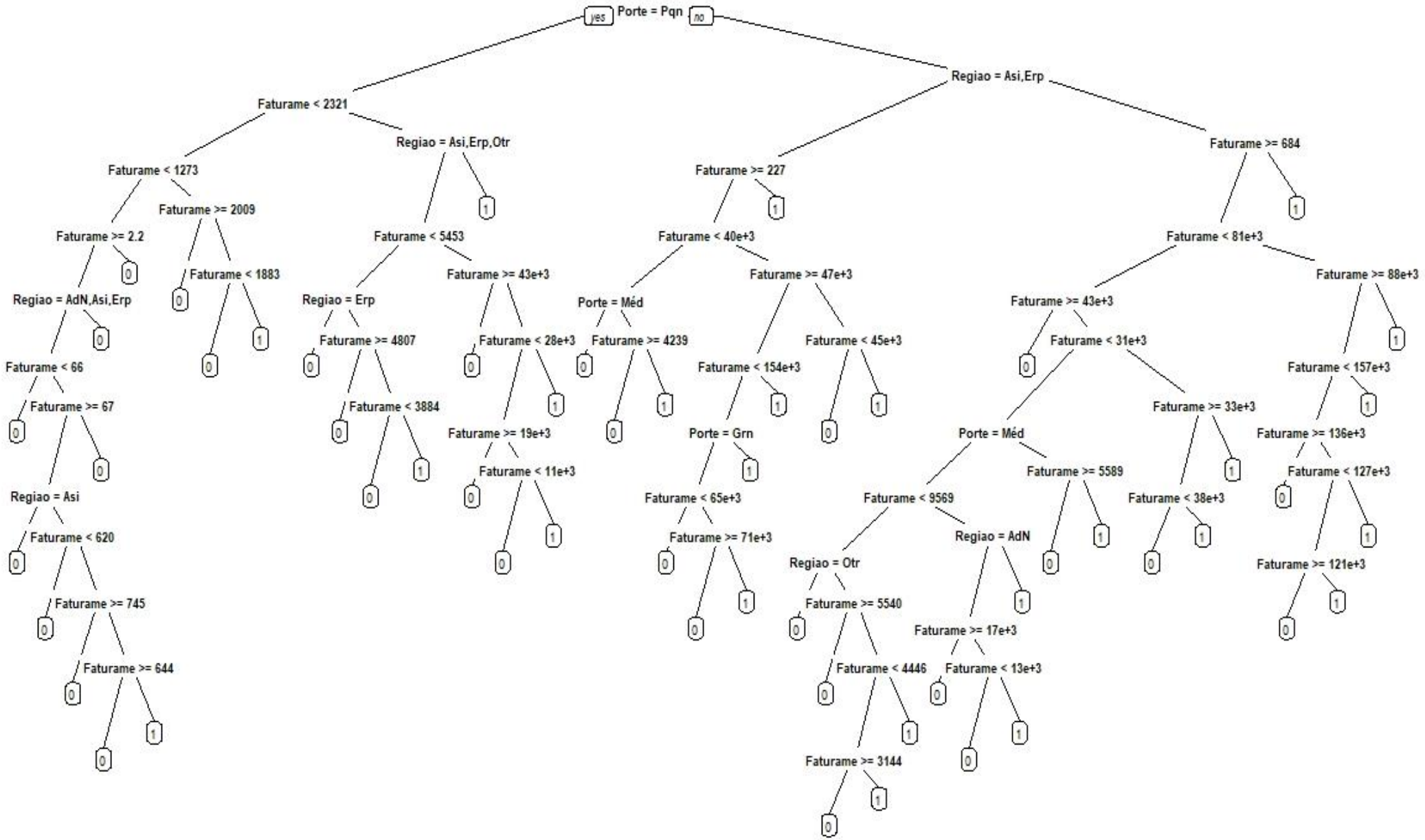

Figura 5. Modelo em árvore de decisão para interpretação.
Fonte: Elaboração própria.

5 CONSIDERAÇÕES FINAIS

O presente estudo destaca o potencial da técnica de indução de árvores de decisão para classificar casos de ocorrência e não ocorrência de sinistros cibernéticos em empresas do setor bancário e identificar a importância das variáveis associadas a partir de um banco de dados de alcance mundial. Os resultados da classificação usando árvore de decisão treinada com algoritmo apresentaram 89% de acertos globais. Utilizando as informações cadastrais de empresas, as seguradoras têm condições de diagnosticar novas empresas em relação a possibilidade de ocorrência de ataque cibernético.

Além disso, quando comparada com a modelagem tradicional GLM, a metodologia de árvores de decisão proporcionou um melhor ajuste ao se verificar um percentual de erro de 11%, metade do percentual alcançado pelo GLM. Em relação à sensibilidade, a diferença percentual foi ainda maior. Enquanto a árvore de decisão classificou corretamente 62% das empresas, o GLM atingiu apenas o percentual de 28%.

Nesse sentido, entende-se que a modelagem em questão garante uma boa qualidade de classificação em relação aos dados utilizados, permitindo que os valores por ela apresentados atuem de forma inovadora como ferramenta de apoio à tomada de decisão das seguradoras, visando respostas úteis ao gerenciamento de riscos cibernéticos. Uma parte importante do processo de precificação é identificar todas as características que permitam prever a quantidade de indenizações futuras e bem selecionar os segurados, cobrando prêmios mais baixos dos grupos de menor risco e mais altos dos grupos de maior risco. Visto que são escassos os estudos que englobam a problemática aqui apresentada, espera-se que o presente estudo fomente maiores discussões acerca da correta estimação do risco que pode ser potencialmente prejudicial economicamente.

REFERÊNCIAS

Allianz. (2022). *11º Allianz Risk Barometer 2022*. https://www.abtra.org.br/inovacao-e-tecnologia/11o-allianz-risk-barometer-2022/

Biener, C., Eling, M., & Wirfs, J. H. (2015). Insurability of cyber risk: An empirical analysis. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 40, 131-158. https://doi.org/10.1057/gpp.2014.19

Bonini, J. A. (2016). Aplicação de algoritmos de árvore de decisão sobre uma base de dados de câncer de mama. *Revista ComInG-Communications and Innovations Gazette*, 1(1), 57-67. https://doi.org/10.5902/2448190421132

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. New York: Routledge.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2), 261-304. https://doi.org/10.1177/0049124104268644

Carfora, M., Martinelli, F., Mercaldo, F., & Orlando, A. (2019). Cyber risk management: An actuarial point of view. *Journal of Operational Risk*, 14(4). https://doi.org/10.21314/JOP.2019.231

ClearSale (2022). *Mapa da fraude 2022.* https://br.clear.sale/mapa-da-fraude

CRO. (2016). *Forum concept paper on a proposed categorisation methodology for cyber risk*. https://www.thecroforum.org

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In M. Cord & P. Cunningham (Eds.), *Machine learning techniques for multimedia* (pp. 21-49). Berlin: Springer.

Dal Moro, E. (2020). Towards an economic cyber loss index for parametric cover based on IT security indicator: A preliminary analysis. *Risks*, 8(2), 45. https://doi.org/10.3390/risks8020045

Ecotrust. (2023). *Gestão de vulnerabilidades na área financeira: por que se preocupar?* https://blog.ecoit.com.br/gestao-de-vulnerabilidades-na-area-financeira/

Eling, M., & Schnell, W. (2016). What do we know about cyber risk and cyber risk insurance?. *The Journal of Risk Finance*, 17(5), 474-491. https://doi.org/10.1108/JRF-09-2016-0122

Faceli, K., Lorena, A. C., Gama, J., & Carvalho, A. C. P. D. L. F. D. (2011). *Inteligência artificial: uma abordagem de aprendizado de máquina.* Rio de Janeiro: LTC.

Fortinet. (2022). *Brasil é o segundo país que mais sofre ataques cibernéticos na América Latina.* https://www.fortinet.com/br/corporate/about-us/newsroom/press-releases/2022/brasil-e-o-segundo-pais-que-mais-sofre-ataques-ciberneticos-na-a

Gai, K., Qiu, M., & Elnagdy, S. A. (2016, April). Security-aware information classifications using supervised learning for cloud-based cyber risk management in financial big data. In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS),* 197-202. New York, United States of America.

Hamoud, A., Hashim, A. S., & Awadh, W. A. (2018). Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5, 26-31. https://doi.org/ 10.9781/ijimai.2018.02.004

Karam, E. (2014). *Measuring and managing operational risk in the insurance and banking sectors* (Doctoral dissertation, Université Claude Bernard-Lyon I).

Lemos, E. P., Steiner, M. T. A., & Nievola, J. C. (2005). Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. *Revista de Administração-RAUSP*, 40(3), 225-234.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522), 626-636. https://doi.org/10.1080/01621459.2016.1264957

Marotta, A., Martinelli, F., Nanni, S., Orlando, A., & Yautsiukhin, A. (2017). Cyber-insurance survey. *Computer Science Review*, 24, 35-61. https://doi.org/10.1016/j.cosrev.2017.01.001

Martinez, E. Z., Louzada-Neto, F., & Pereira, B. D. B. (2003). A curva ROC para testes diagnósticos. *Caderno saúde coletiva*, 11, 7-31.

Pekár, S., & Brabec, M. (2018). Generalized estimating equations: A pragmatic and flexible approach to the marginal GLM modelling of correlated data in the behavioural sciences. *Ethology*, 124(2), 86-93. https://doi.org/10.1111/eth.12713

Peng, C., Xu, M., Xu, S., & Hu, T. (2018). Modeling multivariate cybersecurity risks. *Journal of Applied Statistics*, 45(15), 2718-2740. https://doi.org/10.1080/02664763.2018.1436701

Podhorská, I., Vrbka, J., Lazaroiu, G., & Kovacova, M. (2020). Innovations in financial management: Recursive prediction model based on decision trees. *Marketing and Management of Innovations*, 3, 276-292. https://doi.org/10.21272/mmi.2020.3-20

Portugal, M. S. (1995). Notas introdutórias sobre o princípio de máxima verossimilhança: Estimação e teste de hipóteses. *DECON/UFRGS*, Porto Alegre, Abril.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. São Francisco: Morgan-Kaufmann.

Sembiring, N. S. B., Sinaga, M. D., Ginting, E., Tahel, F., & Fauzi, M. (2021, September). Predict the Timeliness of Customer Credit Payments at Finance Companies Using a Decision Tree Algorithm. In *2021 9th International Conference on Cyber and IT Service Management (CITSM)*, Bengkulu, Indonesia.

RL, M., & Mishra, A. K. (2022). Measuring financial performance of Indian manufacturing firms: application of decision tree algorithms. *Measuring Business Excellence*, 26(3), 288-307. https://doi.org/10.1108/mbe-05-2020-0073

Ruiz-Maya, L. (1978). Sobre la metodología del Índice de Gini. Universidad Autónoma de Madrid. https://repositorio.uam.es/bitstream/handle/10486/5861/36175_6.pdf?sequence=1

Saha, D., Young, T. M., & Thacker, J. (2023). Predicting firm performance and size using machine learning with a Bayesian perspective. *Machine Learning with Applications*, 11, 100453. https://doi.org/10.1016/j.mlwa.2023.100453

Sousa, A. F., Neto, Silva, J. F. G., & Oliveira, G. N. (2021). Predição de Pagamentos Atrasados Através de Algoritmos Baseados em Árvore de Decisão. *Revista de Engenharia e Pesquisa Aplicada,* 6(5), 1-10. https://doi.org/10.25286/repa.v6i5.1746

Subroto, A., & Apriyana, A. (2019). Cyber risk prediction through social media big data analytics and statistical machine learning. *Journal of Big Data*, 6(1), 1-19. https://doi.org/10.1186/s40537-019-0216-1

Vidhya, A. (2021). *Tree Based Algorithms: A Complete Tutorial from Scratch (in R & Python).* https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/

Xu, M., & Hua, L. (2019). Cybersecurity insurance: Modeling and pricing. *North American Actuarial Journal*, 23(2), 220-249. https://doi.org/10.1080/10920277.2019.1566076

Yuvaraj, N., Chang, V., Gobinathan, B., Pinagapani, A., Kannan, S., Dhiman, G., & Rajan, A. R. (2021). Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. *Computers & Electrical Engineering*, 92. https://doi.org/10.1016/j.compeleceng.2021.107186

CONTEXTUS
REVISTA CONTEMPORÂNEA DE ECONOMIA E GESTÃO.
ISSN 1678-2089
ISSNe 2178-9258
1. Economia, Administração e Contabilidade – Periódico
2. Universidade Federal do Ceará. FEAAC – Faculdade de Economia, Administração, Atuária e Contabilidade

FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE (FEAAC)
Av. da Universidade – 2486, Benfica
CEP 60020-180, Fortaleza-CE

DIRETORIA: Paulo Rogério Faustino Matos
Danielle Augusto Peres

Website: www.periodicos.ufc.br/contextus
E-mail: revistacontextus@ufc.br

A Contextus está classificada no sistema Qualis – Capes como periódico B1, na área de Administração Pública e de Empresas, Ciências Contábeis e Turismo (2013-2016).

A Contextus está de acordo e assina a Declaração de São Francisco sobre a Avaliação de Pesquisas (DORA).

A Contextus é associada à Associação Brasileira de Editores Científicos (ABEC).