UNIVERSIDADE FEDERAL DO CEARÁ

FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE

DEPARTAMENTO DE ADMINISTRAÇÃO

CURSO DE CIÊNCIAS ATUARIAIS

MARIA RITA DA SILVA BARROS

APRENDIZAGEM DE MÁQUINA APLICADA NA PREDIÇÃO DE RISCOS CIBERNÉTICOS

FORTALEZA-CE

2023

MARIA RITA DA SILVA BARROS

APRENDIZAGEM DE MÁQUINA APLICADA NA PREDIÇÃO DE RISCOS CIBERNÉTICOS

Monografia apresentada ao Curso de Ciências Atuariais do Departamento de Administração da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Ciências Atuariais.

Orientador: Prof. Ms. Alana Katielli Nogueira Azevedo.

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

B279a Barros, Maria Rita da Silva.
Aprendizagem de máquina aplicada na predição de riscos cibernéticos / Maria Rita da Silva Barros. – 2023.
39 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Curso de Ciências Atuariais, Fortaleza, 2023. Orientação: Prof. Me. Alana Katielli Nogueira Azevedo.

1. Risco cibernético. 2. Aprendizagem de máquina. 3. Segurança cibernética. I. Título.

CDD 368.01

MARIA RITA DA SILVA BARROS

APRENDIZAGEM DE MÁQUINA APLICADA NA PREDIÇÃO DE RISCOS CIBERNÉTICOS

Monografia apresentada ao Curso de Ciências Atuariais do Departamento de Administração da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Ciências Atuariais.

Aprovada em: 27/11/2023.

BANCA EXAMINADORA

_____
Profª. Ms. Alana Katielli Nogueira Azevedo (Orientador)
Universidade Federal do Ceará (UFC)

_____
Profª. Dra. Alane Siqueira Rocha
Universidade Federal do Ceará (UFC)

_____
Profª. Ms. Luciana Moura Reinaldo
Universidade Federal do Ceará (UFC)

A Ryan.

À minha mãe e aos meus irmãos.

AGRADECIMENTOS

À família Barros, por me fornecer desde sempre o melhor ambiente para que eu pudesse me desenvolver.

Aos meus amigos, de hoje ou não, por terem me apoiado sempre que precisei, em especial aos meus queridos companheiros que estiveram ao meu lado ao longo da graduação. Fico imensamente feliz que todos nós já estejamos colhendo o que plantamos durante esses quatro anos.

Ao Crispiano, melhor professor de matemática que eu poderia ter tido no ensino fundamental.

A todas as minhas chefes ao longo da minha ainda curta vida profissional, Cris, Claudiana e Brenda, pelas oportunidades de crescimento e incentivo à minha criatividade.

A todos os professores do tão querido e respeitável curso de Ciências Atuariais, que são incansáveis ao se tratar da formação de estudantes, profissionais e, acima de tudo, de seres humanos.

À professora Alana, pelo seu brilhantismo enquanto minha orientadora.

A todos os outros trabalhadores, funcionários públicos, concursados ou terceirizados da Universidade Federal do Ceará: que o meu conhecimento possa retornar a estes em forma de acesso à previdência e à saúde.

Aos barulhentos Bhaskara e Bob, por protegerem tão bem o meu lar.

Aos professores participantes da banca examinadora Alane Siqueira Rocha e Luciana Moura Reinaldo pelo tempo, pelas valiosas colaborações e sugestões.

“O risco é uma opção, e não um destino.” (BERNSTEIN, 1997).

RESUMO

O risco cibernético é uma das principais ameaças às empresas modernas. Ataques cibernéticos podem causar danos financeiros significativos, perda de propriedade intelectual e até mesmo interrupção das operações. Esta pesquisa se propôs a explorar a eficácia de diferentes modelos de Aprendizagem de Máquina na predição de risco cibernético. Utilizando dados obtidos do IBGE, considerando um conjunto de 16 covariáveis estatisticamente significativas de 16.725 empresas, a análise comparativa foi realizada entre os modelos de redes neurais, árvore de decisão, GAMLSS, GBM e floresta aleatória. Os resultados revelaram que as redes neurais alcançaram a melhor proporção de concordância, com 86% de acerto na identificação de empresas atacadas. Em contraste, o GBM apresentou uma taxa de concordância de 56%, sendo o menor resultado obtido entre os métodos citados. O GAMLSS, embora não tenha obtido resultados tão promissores quanto as redes neurais, demonstrou capacidade de ajuste dos dados a uma distribuição de probabilidade, destacando-se a distribuição de Poisson Zero Inflada (ZIP), com parâmetros significativos para futuras aplicações em precificação de seguros cibernéticos.

Palavras-chave: risco cibernético; aprendizagem de máquina; segurança cibernética

ABSTRACT

Cyber risk stands as one of the prime threats to modern enterprises. Cyber-attacks can lead to substantial financial damages, loss of intellectual property, and even operational disruptions. This research aimed to explore the effectiveness of different Machine Learning models in predicting cyber risk. Using data obtained from IBGE (Brazilian Institute of Geography and Statistics), comprising 16 statistically significant covariates across 16,725 companies, a comparative analysis was conducted among neural networks, decision tree, GAMLSS, GBM, and random forest models. The findings revealed that neural networks achieved the highest agreement rate, with an 86% accuracy in identifying targeted companies. In contrast, GBM showed an agreement rate of 56%, marking the lowest among the mentioned methods. While GAMLSS didn't yield as promising results as neural networks, it demonstrated the capability to fit the data to a probability distribution, notably the Poisson Zero-Inflated (ZIP) distribution, with significant parameters for potential applications in cyber insurance pricing.

Keywords: cyber risk; machine learning; cyber security

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A sociedade contemporânea é cada vez mais digital e interconectada. Essa realidade tem impulsionado o uso de tecnologias de informação e comunicação (TICs) por organizações de todos os setores. No entanto, essa dependência digital também aumenta a vulnerabilidade a ataques cibernéticos. Ataques desses tipos podem causar danos financeiros significativos, perda de propriedade intelectual e até mesmo interrupção das operações.

Nesse contexto, a segurança cibernética é uma preocupação fundamental para as organizações. É importante avaliar o risco cibernético, que é a probabilidade e o impacto de ataques cibernéticos, intrusões maliciosas e violações de dados. A avaliação do risco cibernético pode ajudar as organizações a tomar medidas para mitigar esse risco.

A aprendizagem de máquina (machine learning) é uma área da inteligência artificial que permite que computadores aprendam sem serem explicitamente programados. A aprendizagem de máquina tem sido usada com sucesso em uma variedade de aplicações, incluindo a predição de risco cibernético. A predição de risco cibernético é um desafio complexo, pois os ataques cibernéticos são geralmente aleatórios e imprevisíveis. No entanto, a aprendizagem de máquina pode ser usada para identificar padrões nos dados que podem ser usados para prever o risco de ataques cibernéticos.

Este estudo tem como objetivo aplicar alguns modelos de aprendizagem de máquina tais como redes neurais, árvore de decisão, GAMLSS, GBM e floresta aleatória, para comparar e verificar qual a melhor alternativa na predição de sinistros cibernéticos.

A importância deste trabalho é notável porque fornece uma comparação dos resultados de diferentes modelos de aprendizagem de máquina na predição de risco cibernético. Os resultados deste trabalho podem ajudar as empresas a selecionar o modelo de aprendizagem de máquina mais adequado para suas necessidades.

Além desta seção, esta monografia tem em sua estrutura um referencial teórico, que aborda a literatura acerca de machine learning e risco cibernético; a metodologia, que apresenta as fontes de dados, as informações coletadas e detalhamento acerca do seu uso; os resultados, no qual são expostos e discutidos os resultados encontrados; e, por fim, as considerações finais, na qual são apresentadas as conclusões deste trabalho.

2 REFERENCIAL TEÓRICO

2.1 Riscos cibernéticos

O cenário atual da sociedade moderna está profundamente impregnado pela digitalização e pela intensa interconexão viabilizada pelas tecnologias de informação e comunicação. Nesse contexto, a segurança cibernética emergiu como uma preocupação fundamental e premente para organizações de todos os setores, em virtude da crescente dependência de sistemas digitais e da constante proliferação de ameaças virtuais. O risco cibernético, portanto, se estabelece como um conceito de extrema relevância, abrangendo a avaliação da probabilidade e do impacto de ataques cibernéticos, intrusões maliciosas e violações de dados. Esses eventos podem acarretar consequências significativas, não apenas do ponto de vista financeiro, mas também em termos de operações e reputação, representando uma ameaça abrangente e multifacetada para as organizações. De acordo com Georg (2023) "os crimes cibernéticos e a insegurança cibernética se encontram entre os dez maiores riscos a serem enfrentados globalmente […] dando a dimensão do desafio global no seu equacionamento." Consequentemente, a salvaguarda da segurança cibernética tornou-se, assim, uma prioridade incontestável diante dos desafios inerentes a um mundo cada vez mais digital e interconectado.

> A perturbação da integridade ou autenticidade de dados ou informações é chamada de ataque de rede de computadores ou ciberataque. O código malicioso que altera a lógica de um programa e causa erros na saída. O processo de hacking envolve a varredura da Internet em busca de sistemas que possuam controles de segurança inadequados e que estejam mal configurados. Uma vez que um hacker infecta o sistema, ele/ela pode operá-lo remotamente e enviar comandos para que o sistema atue como um espião para os invasores, além de ser usado para interromper outros sistemas. O hacker espera que o sistema infectado tenha algumas falhas, como bugs no software, falta de antivírus e configuração deficiente do sistema, de modo que outros sistemas possam ser infectados por meio desse sistema. O ciberataque tem como objetivo roubar ou hackear as informações de qualquer organização ou órgão governamental. Para roubar os dados ou informações, o atacante ou hacker segue certas características para que possam alcançar seus objetivos (UMA; PADMAVATHI, 2013).

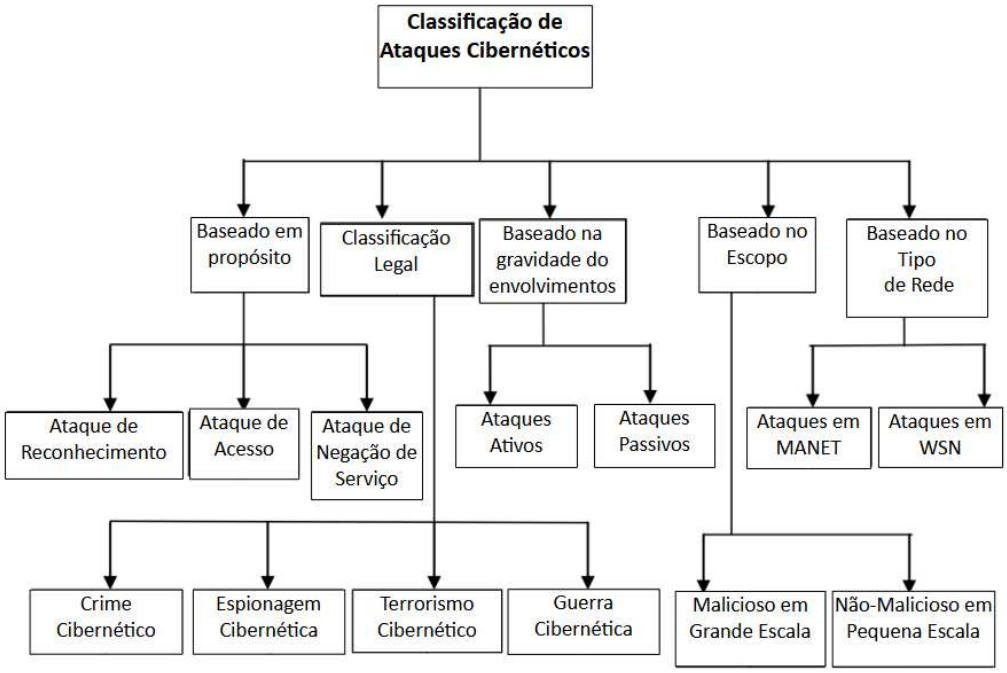
A caracterização de um risco como cibernético é estabelecida quando a ocorrência da perda está intrinsecamente ligada ao uso de programas de computadores e/ou à internet. Além disso, essa perda está diretamente relacionada a danos ou à perda de dados, que podem se manifestar de diversas formas, como modificações não autorizadas, perda de exclusividade (em casos de violações de dados) ou abusos envolvendo informações em geral (BÖHME; LAUBE; RIEK, 2019).

Perdas como as supracitadas implicam em custos. Neste sentido, de acordo com Justo (2018), quando ocorre um incidente cibernético nas empresas, surgem custos imediatos e custos indiretos. Os custos imediatos são inevitáveis e abrangem despesas legais e investigações forenses, custos relacionados à comunicação com os clientes, possíveis interrupções nos negócios, custos de fraude, extorsão e danos físicos. Por outro lado, os custos indiretos variam de acordo com a gravidade do evento e incluem tanto o impacto imediato quanto os custos a longo prazo. Estes envolvem despesas com processos movidos por terceiros, prejuízos à reputação, multas ou penalidades regulatórias, impacto no valor das ações, perdas na vantagem competitiva e, consequentemente, prejuízos futuros na receita.

A complexidade dos riscos cibernéticos não se limita a essa caracterização fundamental, pois eles podem ser classificados com base em vários critérios adicionais. Um desses critérios é a atividade envolvida, que pode ser classificada como criminal ou não-criminal. Os ataques cibernéticos também podem ser categorizados de acordo com o tipo de método utilizado, abrangendo categorias como malware, ataques internos, spam, negação de serviço distribuída e muitos outros. Adicionalmente, a origem dos ataques é um fator relevante a ser considerado, visto que eles podem ser perpetrados por indivíduos ou entidades de diferentes naturezas, incluindo terroristas, criminosos comuns e até mesmo governos (ELING; SCHNELL, 2016).

Essa abordagem mais abrangente e detalhada da classificação e definição dos riscos cibernéticos reconhece a natureza multifacetada e em constante evolução das ameaças digitais na era da tecnologia. Tais ameaças exigem uma compreensão aprofundada e uma abordagem proativa para a segurança cibernética, à medida que organizações e indivíduos buscam proteger-se contra uma gama diversificada de desafios e cenários potencialmente prejudiciais.

Figura 1 – Classificação de Ataques Cibernéticos



Fonte: (UMA; PADMAVATHI, p. 392, 2013)

De acordo com Bendovschi (2015), diversas definições dos termos cyberattack e cybercrime podem ser encontradas na literatura internacional, todas com o objetivo comum de comprometer a confidencialidade, integridade e disponibilidade de dados. A evolução tecnológica também traz consigo o avanço do cibercrime, resultando na criação contínua de novas maneiras de realizar ataques, atingir alvos cada vez mais difíceis de penetrar e permanecer não rastreados. No entanto, ameaças cibernéticas tradicionais ainda são a fonte dos ataques mais comuns.

À medida que os avanços tecnológicos continuam a moldar nosso mundo, a demanda por segurança no ambiente virtual se torna cada vez mais premente. Os riscos cibernéticos, intrinsecamente ligados à era digital, são agora objeto de estudo e gestão no campo da cibersegurança, uma disciplina dedicada à proteção contra ameaças online. Nesse contexto, surge uma necessidade premente de compreender, avaliar e mitigar os riscos cibernéticos. O mercado de seguros cibernéticos, apesar de seu crescimento, ainda se encontra em um estágio de maturidade inicial e enfrenta uma série de desafios significativos. Estes incluem a identificação precisa dos riscos, a sua avaliação e a especificação adequada da cobertura de seguros (MAROTTA et al., 2017). Além disso, a determinação dos prêmios de seguro cibernético apresenta-se como um desafio complexo no cenário atual.

Embora exista uma base teórica substancial sobre o seguro cibernético, há uma notável carência de informações práticas disponíveis publicamente sobre o conteúdo específico das apólices e sobre os métodos empregados pelas seguradoras para calcular os prêmios de seguros cibernéticos (ROMANOSKY et al., 2019). Nesse sentido, torna-se evidente a importância de desenvolver modelos precisos e abrangentes para a previsão de riscos cibernéticos.

Além disso, à medida que a segurança cibernética assume um papel crucial no mundo contemporâneo, a pesquisa e o desenvolvimento de ferramentas, modelos e práticas que melhorem a compreensão e a gestão dos riscos cibernéticos tornam-se essenciais. A criação de modelos concisos para a predição de riscos cibernéticos representa um passo fundamental na direção de uma maior resiliência e proteção nos ambientes digitais em constante evolução.

Entre as várias abordagens para a modelagem de riscos cibernéticos, Cebula e Young (2010) enfatizam a importância de uma organização estruturada das informações disponíveis em quatro classes distintas, proporcionando uma compreensão holística dos desafios associados à segurança cibernética. Essas quatro categorias incluem ações desencadeadas por pessoas, falhas em sistemas e tecnologias, processos internos suscetíveis a falhas e eventos externos. A categorização abrangente das ameaças cibernéticas em tais grupos fornece uma base sólida para a análise e a mitigação de riscos em ambientes digitais cada vez mais complexos e interconectados.

Além disso, a Aprendizagem de Máquina surge como uma ferramenta de extrema relevância na modelagem e previsão de riscos cibernéticos, com aplicações abrangentes em diversos domínios críticos. Essa abordagem demonstrou sua eficácia em cenários tão diversos quanto o reforço da segurança em usinas elétricas, a detecção ágil de ataques cibernéticos em sistemas de controle industrial por meio de divisões de zona, a identificação de invasões em sistemas SCADA (Supervisory Control and Data Acquisition), a detecção de intrusões em redes veiculares (VANETs - Vehicular Ad Hoc Networks) e a análise de malwares (HANDA; SHARMA; SHUKLA, 2019).

O potencial transformador da Aprendizagem de Máquina no tratamento dos riscos cibernéticos reflete sua capacidade ímpar de processar volumes massivos de dados, identificar padrões e tendências, e adaptar-se dinamicamente a ameaças que estão em constante evolução. Essa abordagem promissora desempenha um papel crucial na salvaguarda de ativos digitais e na garantia da segurança cibernética em um mundo cada vez mais interdependente da tecnologia. Conforme novos desafios e ameaças emergem, a aplicação da Aprendizagem de Máquina continua a expandir seu alcance, fortalecendo nossa capacidade de enfrentar os riscos

cibernéticos de forma proativa e eficaz, alinhando-se com a evolução constante do cenário de segurança cibernética.

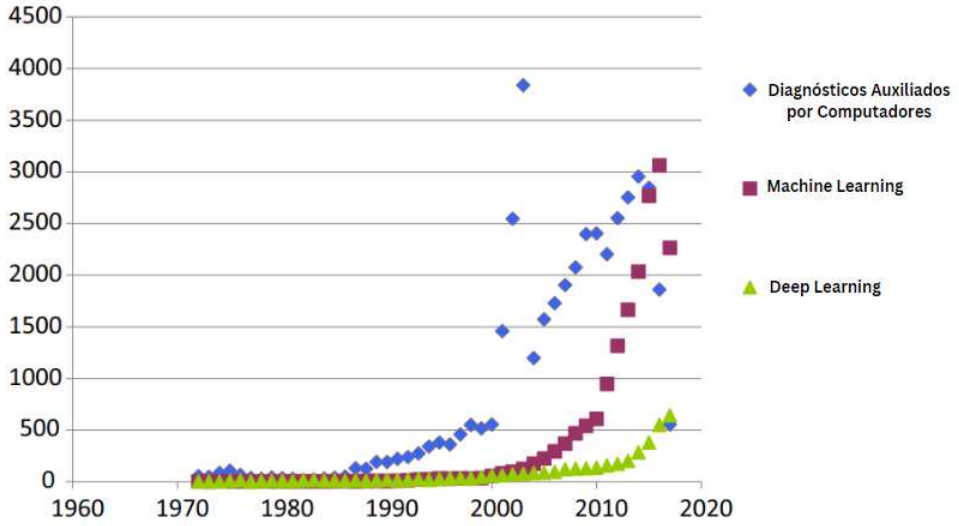
2.2 Aprendizagem de máquina

Nos últimos anos, o campo de Machine Learning (Aprendizagem de Máquina) tem vivenciado um crescimento extraordinário, impulsionado por diversos fatores que convergem para criar um ambiente propício ao desenvolvimento e à aplicação de algoritmos cada vez mais sofisticados. A disponibilidade crescente de dados complexos, juntamente a avanços tecnológicos notáveis, tem desempenhado um papel fundamental nesse cenário. Esta revolução no campo da Aprendizagem de Máquina tem proporcionado inúmeras oportunidades para extrair conhecimento significativo a partir de grandes conjuntos de dados em uma variedade de domínios, desde a medicina até as análises de mercado.

Com a proliferação de dispositivos conectados à internet e a coleta de dados em escala sem precedentes, a capacidade de aproveitar os recursos do Machine Learning tornou-se fundamental. Diagnósticos médicos são aprimorados por meio de algoritmos de Machine Learning que podem analisar vastos conjuntos de dados médicos e históricos de pacientes para identificar padrões e tendências que seriam difíceis, se não impossíveis, de se detectar por métodos tradicionais (RAJKOMAR; DEAN; KOHANE, 2019). No mundo dos negócios e do mercado financeiro, as análises de mercado se beneficiam do Machine Learning para entender o comportamento do consumidor, otimizar estratégias de preços, prever tendências e melhorar a tomada de decisões (WHIG, 2021).

Giger (2018) elabora um estudo sobre as pesquisas científicas da área médica e cataloga artigos publicados na revista científica PubMed, uma das mais renomadas nos estudos de medicina, a respeito de Machine Learning, Deep Learning e Diagnósticos Auxiliados por Computadores em Radiologia num período de tempo de 1972 a meados de 2017.

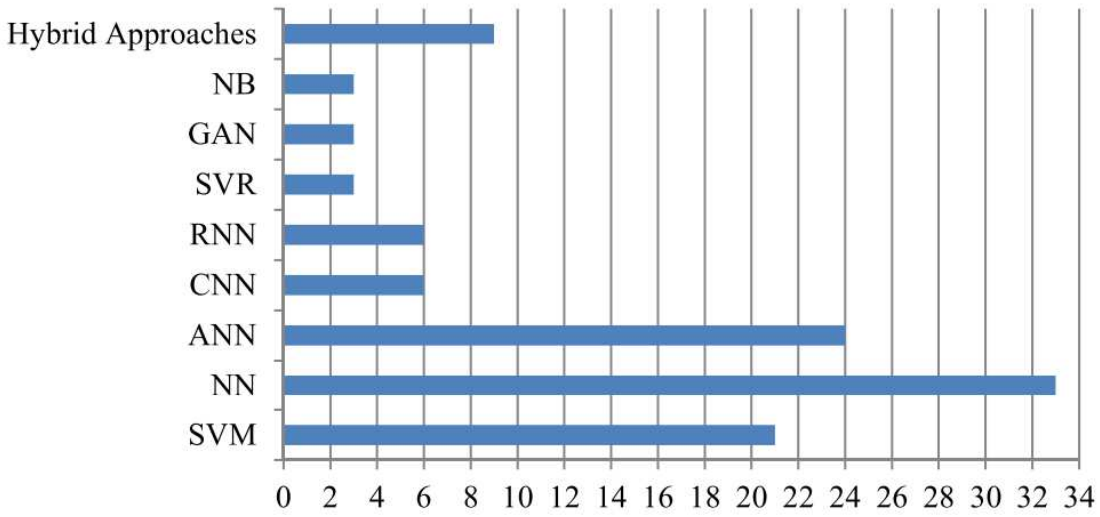
Gráfico 1 – Número de Papers contabilizados na PubMed



Fonte: Traduzida de (GIGER, p. 512, 2018).

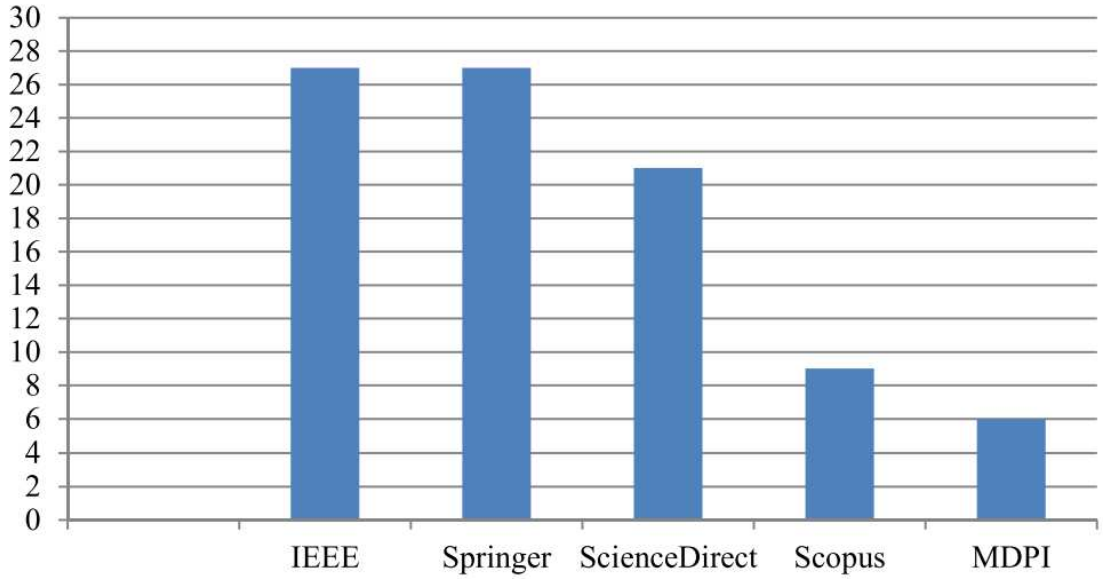
Já no que diz respeito ao mercado financeiro, mais especificamente às predições possíveis no mercado de ações, Kumar, Sarangi e Verma (2022) expõem em sua revisão sistemática não só os percentuais das técnicas mais frequentes de ML utilizadas para estas predições como também das revistas científicas que mais publicam artigos sobre a temática em questão.

Gráfico 2 – Técnicas mais frequentes para predições no mercado de ações



Fonte: Adaptada de (KUMAR; SARANGI; VERMA, p. 4, 2022).

Gráfico 3 – Revistas com maior frequência de estudos sobre ML no mercado de ações



Fonte: (KUMAR; SARANGI; VERMA, p. 5, 2022).

Entretanto, um domínio que merece atenção especial é o da segurança cibernética. Com a crescente complexidade das ameaças digitais e a diversificação das táticas utilizadas por cibercriminosos, a necessidade de técnicas avançadas de Machine Learning se tornou ainda mais premente. A capacidade de analisar rapidamente enormes volumes de dados em busca de anomalias, identificar potenciais ameaças e, o mais importante, aprender com essas ameaças em evolução constante, é essencial para a defesa cibernética eficaz.

Neste contexto, os avanços recentes em técnicas de Machine Learning têm sido um farol de esperança. Algoritmos de aprendizado profundo, redes neurais convolucionais e redes neurais recorrentes têm mostrado um potencial notável na detecção de ameaças cibernéticas. Esses modelos são capazes de analisar padrões de tráfego de rede, identificar comportamentos suspeitos e até mesmo prever ataques antes que ocorram (MENG, 2011). A análise de logs de servidores, registros de eventos de segurança e outras fontes de dados se beneficia do Machine Learning, permitindo uma resposta mais rápida e eficiente às ameaças em um cenário de segurança cibernética em constante mudança.

À medida que o campo da Aprendizagem de Máquina continua a evoluir, podemos esperar avanços ainda mais notáveis na aplicação dessa tecnologia em domínios críticos, como a segurança cibernética. A capacidade de lidar com dados complexos e em constante evolução é uma das principais vantagens do Machine Learning, tornando-o uma ferramenta indispensável na nossa busca contínua por insights valiosos e na proteção contra ameaças digitais cada vez mais sofisticadas.

A ciência da Aprendizagem de Máquina (Machine Learning) concentra-se na pesquisa e desenvolvimento de algoritmos e modelos estatísticos capazes de capacitar sistemas computacionais a realizar tarefas específicas sem a necessidade de uma programação explícita por parte dos humanos. Essa área representa um campo de estudo em constante evolução que tem desencadeado transformações significativas em diversos setores. Uma das características mais notáveis da aprendizagem de máquina é a sua capacidade de automação uma vez que um algoritmo tenha aprendido como lidar com um conjunto de dados específico (MAHESH, 2020).

Neste contexto em constante evolução, a Aprendizagem de Máquina desempenha um papel cada vez mais significativo na automação de tarefas complexas, na análise de dados em larga escala e na resolução de problemas em diversas disciplinas, desde medicina até a indústria financeira. À medida que a pesquisa e o desenvolvimento continuam, podemos esperar avanços adicionais na capacidade das máquinas de aprender e aplicar conhecimento, com implicações profundas em nosso mundo tecnológico em constante mudança.

O cerne da aprendizagem de máquina gira em torno da criação de algoritmos de aprendizado que são capazes de construir modelos a partir de dados adquiridos por meio de experiências passadas. Ao fornecer a esses algoritmos conjuntos de dados que representam experiências anteriores, é possível treiná-los para desenvolver modelos que, por sua vez, têm a capacidade de realizar previsões e tomadas de decisão em relação a novas observações. Essa capacidade de generalização, em que os modelos podem aplicar o conhecimento adquirido a situações inéditas, é uma das características mais poderosas da aprendizagem de máquina. Isso implica que, à medida que um algoritmo é exposto a mais dados e experiências, ele pode aprimorar e refinar seus modelos, tornando-se mais preciso e eficiente na realização das tarefas específicas para as quais foi treinado. (ZHOU, 2021).

Uma tendência promissora e notável na evolução da Aprendizagem de Máquina é a adoção generalizada de modelos de Aprendizado Profundo (Deep Learning), uma subcategoria do campo mais amplo do aprendizado de máquina. O Aprendizado Profundo se destaca por sua arquitetura complexa, caracterizada por um número significativo de camadas e parâmetros em comparação com abordagens mais tradicionais (SHINDE; SHAH, 2018).

Para compreender o funcionamento desses modelos, é fundamental ter um vislumbre do que é uma Rede Neural, que é a base desses sistemas. Uma Rede Neural é um modelo de computação que se assemelha à estrutura de neurônios no cérebro humano. Ela consiste em uma vasta rede de nós interconectados, que podem ser comparados aos neurônios no cérebro. Cada nó desempenha um papel específico ao aplicar uma função de output particular, conhecida como função de ativação. As conexões entre quaisquer dois nós têm associados pesos

que representam a força do sinal transmitido através da ligação, desempenhando um papel análogo à memória da rede neural artificial (WU; FENG, 2018).

Essa arquitetura de múltiplas camadas e conexões ponderadas é a essência do Aprendizado Profundo, permitindo a criação de modelos altamente complexos capazes de aprender e representar informações em níveis abstratos e hierárquicos. À medida que mais camadas são adicionadas, a rede pode extrair características cada vez mais sofisticadas e relevantes dos dados, tornando-a particularmente adequada para tarefas que envolvem dados complexos, como reconhecimento de padrões, processamento de linguagem natural e visão computacional. A expansão do Aprendizado Profundo representa uma área de pesquisa e aplicação fascinante que está moldando o futuro da tecnologia e tem potencial para impactar positivamente uma ampla gama de setores (LECUN; BENGIO; HINTON, 2015).

Outro aspecto de aplicação amplamente promissor da Aprendizagem de Máquina reside na área de segurança cibernética, notadamente na detecção de intrusões e anomalias de rede. A segurança cibernética é uma preocupação crítica em um mundo cada vez mais digital e interconectado, e a detecção de atividades suspeitas desempenha um papel fundamental na proteção de sistemas e dados. A questão central na detecção de anomalias envolve a abordagem de classificação, que se concentra na tarefa de distinguir de forma eficaz e eficiente entre atividades consideradas normais e aquelas que são classificadas como anormais ou potencialmente prejudiciais.

Para abordar esse desafio complexo, os métodos de aprendizado de máquina têm se destacado como uma abordagem poderosa e em constante evolução. A comunidade de detecção de intrusões tem investido significativamente no desenvolvimento e aprimoramento de algoritmos de aprendizado de máquina, com o objetivo de aprimorar o desempenho na detecção de anomalias (MENG, 2011). Esses algoritmos podem analisar o tráfego de rede em tempo real, identificar padrões suspeitos, e tomar medidas preventivas para mitigar ameaças em potencial.

Em resumo, a aplicação da Aprendizagem de Máquina na detecção de intrusões e na segurança cibernética representa uma frente crítica na defesa contra ameaças digitais. À medida que os sistemas de aprendizado de máquina continuam a evoluir, podemos esperar um aprimoramento constante na capacidade de detectar e mitigar ameaças, tornando nossos ambientes digitais mais seguros e resilientes.

3 PROCEDIMENTOS METODOLÓGICOS

3.1 Predição com o uso de redes neurais

Uma Rede Neural é um modelo de gerenciamento de informações que se inspira no funcionamento dos sistemas nervosos biológicos do cérebro humano. Uma grande vantagem da aplicação de Redes Neurais Artificiais (ANNs) é que elas podem tornar modelos mais fáceis de usar e mais precisos em sistemas naturais complexos com entradas extensas. As ANNs são consideradas um modelo muito inovador e útil quando aplicadas à resolução de problemas e Machine Learning (ABIODUN et al., 2018).

Matematicamente, um neurônio artificial i recebe inputs X, ponderados cada um por um peso W. O somatório destas variáveis, sendo adicionado de um termo de viés W_0, são processados numa função f(s) (função de ativação). De acordo com Yadav, Yadav e Kumar (2015) o output de um neurônio artificial i, portanto, é dado por:

$$\boldsymbol{Oi = f\left[w_0 + \sum_{j=1}^{n} w_{ij} x_{ij}\right]} \quad (1)$$

3.2 Predição com o uso de árvore de decisão

As árvores de decisão (DTs) são um dos modelos de inteligência computacional (CI) mais amplamente utilizados. Mesmo quando outros algoritmos fornecem modelos mais precisos (que se aproximam melhor do alvo), as DTs são frequentemente consideradas muito atraentes. Uma das razões mais importantes para sua atratividade é a facilidade de compreensão. As DTs podem ser facilmente expressas na forma de um conjunto de regras lógicas que descrevem as funções de decisão. Quando usadas para suporte à decisão, as DTs fornecem explicações simples para decisões específicas, geralmente na forma de uma única regra lógica (que se aplica ao caso em questão) sendo uma conjunção de várias premissas legíveis (GRĄBCZEWSKI, 2014).

Dentre os algoritmos de árvore de decisão (ID3, C4.5, C5.0, CART), destacam-se ID3 (Iterative Dichotomiser 3) e C4.5 como os mais utilizados (ANURADHA e VELMURUGAN, 2014).

Os algoritmos ID3 e C4.5 são algoritmos baseados em ganho de informação desenvolvidos por Ross Quinlan. O algoritmo ID3 constrói árvores de decisão com base no ganho de informação obtido a partir dos dados de treinamento, enquanto o C4.5 utiliza uma informação adicional chamada taxa de ganho. A árvore de decisão construída é então usada para

classificar os dados de teste. O conjunto de dados projetado para treinar a árvore de decisão serve como entrada para os algoritmos e é composto por objetos e várias características. O processo de construção da árvore de decisão para ambos os algoritmos é semelhante.

Primeiramente, a entropia da classe e a entropia de cada atributo são calculadas, em seguida, o ganho de informação é calculado para todos os atributos, conforme visto nas equações I, II e III abaixo. No algoritmo ID3, o atributo com o maior ganho de informação é considerado o atributo mais informativo e é selecionado como o nó raiz. E o processo é repetido até que todos os atributos estejam na árvore (MIENYE, 2019).

(I) $$info(D) = -\sum_{i=1}^{m} pi * log_2(pi) \quad (2)$$

(II) $$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{D} * info(D_j) \quad (3)$$

(III) $$Ganho(A) = info(D) - info_A(D) \quad (4)$$

3.3 Predição com o uso de GAMLSS

Os Modelos Generalizados Aditivos para Localização, Escala e Forma (GAMLSS) foram introduzidos por Rigby e Stasinopoulos em 2005. Os GAMLSS são definidos como modelos semi-paramétricos. Na verdade, além de exigir uma definição de uma distribuição paramétrica para a variável de resposta, é possível adicionar funções de suavização não paramétricas para cada parâmetro considerado na especificação do modelo. Os autores apresentaram os GAMLSS como uma maneira de superar algumas limitações dos Modelos Lineares Generalizados (GLM) e dos Modelos Aditivos Generalizados (GAM).

Os GLM foram introduzidos por Nelder e Wedderburn em 1972 e representam uma generalização do modelo de regressão linear, no qual é possível usar como variável de resposta uma distribuição de probabilidade diferente da normal. Uma generalização adicional é representada pelos GAM, introduzidos por Hastie e Tibshirani em 1990, como uma extensão dos GLM, onde um componente de suavização não paramétrico é considerado.

Em comparação com os GLM e GAM, as características básicas dos GAMLSS são duas. Em primeiro lugar, a suposição de distribuição da família exponencial para a variável de resposta é substituída por uma família de distribuição mais geral.

Além disso, os GAMLSS permitem expandir a modelagem para parâmetros de escala e forma, como assimetria e curtose. Por essas razões, eles são particularmente flexíveis e adequados para modelar dados nos quais a variável de resposta apresenta algumas dessas características (MARLETTA, 2021).

Ainda de acordo com Marletta (2021), a formulação original é GAMLSS é dada por:

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} Z_{jk}\gamma_{jk} \quad (5)$$

Onde para k = 1, 2, 3, 4, $g_k(.)$ são funções de link que relacionam a distribuição de parâmetros às variáveis explanatórias, X_k é uma matriz *design* conhecida de ordem $n \times J'_k$, $\beta'_k = (\beta_1, \ldots, \beta_{J'_k})$ é um vetor paramétrico de tamanho J'_k e $Z_{jk}\gamma_{jk}$ os termos não-paramétricos adicionais.

3.4 Predição com o uso de GBM

O Gradient Boosting Machine (GBM) é uma técnica poderosa e amplamente utilizada em aprendizado de máquina. Ela alcança resultados de ponta em uma variedade de tarefas, incluindo classificação, regressão e ordenação. O GBM tem várias implementações, como o XGBoost, LightGBM e CatBoost. Em essência, o GBM combina previsões básicas de maneira gananciosa (greedy), procurando as melhores divisões e expandindo uma árvore única de forma gananciosa. Introduzir variabilidade artificial adicional nessas etapas é crucial para melhorar o desempenho do algoritmo (CHANG, 2019).

"Gradient" se refere ao erro ou resíduo obtido após a construção de um modelo, enquanto "boosting" se refere a aprimorar. A técnica é chamada de máquina de gradient boosting, ou GBM. O gradient boosting é uma maneira de gradualmente aprimorar (reduzir) o erro (AYYADEVARA, 2018).

Ayyadevara (2018) também descreve um passo-a-passo da aplicação do GBM na qual se aprimora uma Árvore de Decisão:

I. Inicializa-se uma predição com uma árvore de decisão simples;

II. Calcula-se o resíduo (previsão real do valor);

III. Constrói-se outra árvore de decisão simples que prediga resíduos com base em todos os valores independentes;

IV. Atualize a predição original com a nova predição multiplicada pela taxa de aprendizagem;

V. Repita os passos 2 a 4 para um certo número de iterações (o número de iterações será o número de árvores).

3.5 Predição com o uso de floresta aleatória

Floresta Aleatória é um procedimento de Machine Learning popular que pode ser

utilizado para desenvolver modelos de previsão. Primeiramente introduzida por Breiman em 2001 (BREIMAN, 2001). as Florestas Aleatórias consistem em uma coleção de árvores de decisão para classificação e regressão. As árvores de decisão são fáceis de usar na prática e oferecem um método intuitivo para prever resultados, diferenciando valores "altos" e "baixos" de uma variável preditora em relação ao resultado. No entanto, a metodologia de árvore de decisão muitas vezes apresenta baixa precisão para conjuntos de dados complexos, como conjuntos de dados grandes ou conjuntos de dados com interações complexas entre variáveis.

No contexto da Floresta Aleatória, várias árvores de classificação e regressão são construídas usando conjuntos de treinamento selecionados aleatoriamente e subconjuntos aleatórios de variáveis preditoras para modelar os resultados. Os resultados de cada árvore são agregados para fornecer uma previsão para cada observação. Portanto, a Floresta Aleatória frequentemente oferece maior precisão em comparação com um único modelo de árvore de decisão, mantendo algumas das qualidades benéficas dos modelos de árvore (SPEISER et al, 2019).

Um esquema fundamental para analisar as propriedades teóricas das florestas envolve modelos nos quais as divisões não dependem do conjunto de treinamento D_n. Essa categoria simplificada de modelos é frequentemente chamada de "florestas puramente aleatórias," nas quais o espaço de entrada é representado por X= 〚[0,1]〛^d.

Um exemplo comum é a "floresta central," que opera da seguinte maneira: (i) não há reamostragem de dados; (ii) em cada nó de cada árvore individual, é selecionada aleatoriamente uma coordenada a partir do conjunto {1,…,p}; e (iii) uma divisão é feita no centro da célula ao longo da coordenada escolhida. Essas operações (ii) a (iii) são repetidas k vezes de forma recursiva, onde k ∈N é um parâmetro do algoritmo. O processo é encerrado quando uma árvore binária completa com k níveis é atingida, de modo que cada árvore tem precisamente 2^k folhas. A estimativa final no ponto de consulta x é obtida calculando a média dos Y_icorrespondentes aos X_i que se encontram na célula que inclui x (BIAU e SCORNET, 2016).

3.6 Medidas de desempenho

Uma das formas de representação para verificar o desempenho dos modelos de aprendizagem de máquina 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 6, 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 7).

$$err = \frac{FP+FN}{VP+FP+VN+FN} \tag{6}$$

$$ac = \frac{VP+VN}{VP+FP+VN+FN} \tag{7}$$

De acordo com Martinez et al. (2003), a sensibilidade (Equação 8) 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 9) é a probabilidade de o teste fornecer resultado negativo, traduzindo a capacidade do teste de identificar uma empresa que não sofre sinistro cibernético.

$$sens = \frac{VP}{VP+FN} \tag{8}$$

$$esp = \frac{VN}{VN+FP} \tag{9}$$

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

4.1 Base de dados

O Instituto Brasileiro de Geografia e Estatística (IBGE) é o principal fornecedor de dados e informações do país, atendendo aos mais diversos segmentos da sociedade civil, bem como aos órgãos governamentais federais, estaduais e municipais.

No ano de 2010, a pesquisa sobre o uso das Tecnologias de Informação e Comunicação (TIC) investigou aspectos do uso dessas tecnologias pelo segmento empresarial brasileiro. Entre seus temas, a pesquisa trouxe informações sobre o uso de computadores e da internet nas atividades dessas organizações e os motivos apresentados para explicar o seu não uso. Também foram apresentadas informações sobre as políticas de segurança de TIC adotadas e as competências do pessoal empregado em relação a essas tecnologias.

Os dados utilizados na pesquisa foram formatados em duas etapas. Todas as informações foram anonimizadas e desidentificadas pelo IBGE antes da análise. A primeira etapa envolveu selecionar apenas as informações de empresas que utilizavam computadores e internet. Foram consideradas 16.725 empresas. Uma vez constituídas as empresas selecionadas, a segunda etapa envolveu a escolha das covariáveis estatisticamente significativas para que se pudesse desenvolver o estudo proposto.

A Tabela 2 apresenta as características das 16 covariáveis explicativas escolhidas na pesquisa sobre cada empresa, todas binárias. A Tabela informa, por exemplo, que 66,1% das empresas tinham *homepage,* 61,5% realizavam compras através da internet e 41,8% utilizaram banda larga móvel como conexão à internet. Para o presente estudo, importante ressaltar que 58,8% das empresas sofreram incidentes cibernéticos.

Tabela 2 – Covariáveis explicativas

DESCRIÇÃO	VARIÁVEL	MÉDIA	DESVIO PADRÃO
A empresa possui departamento (área) de TI	dep_ti	0.583	0.493
A empresa proporcionou treinamento/qualificação ao seu pessoal para desenvolver ou aperfeiçoar as habilidades em TI/TIC em 2010	quali_ti	0.445	0.497
A empresa tinha uma política de segurança em TIC formalmente definida, em 2010	seg_ti	0.364	0.481
A empresa dispunha de rede local (LAN) COM fio em dezembro de 2010	lan_f	0.835	0.371
A empresa dispunha de rede local (LAN) SEM fio em dezembro de 2010	lan_sf	0.553	0.497
A empresa dispunha de Intranet em dezembro de 2010	intranet	0.399	0.489
A empresa dispunha de Extranet em dezembro de 2010	extranet	0.271	0.444
A empresa utilizou computação em nuvem (cloud computing) em dezembro de 2010	nuvem	0.209	0.406
Os softwares utilizados pela empresa em 2010, eram prontos para uso (pacotes)	soft_pronto	0.954	0.209
Os softwares utilizados pela empresa em 2010, eram softwares livres	soft_livre	0.585	0.493
Os softwares utilizados pela empresa em 2010, eram desenvolvidos por outra empresa ou através de consultoria	soft_consult	0.753	0.431
A empresa tinha uma página ou portal na internet em dezembro de 2010	homepage	0.661	0.473
A empresa utilizou banda larga fixa como conexão à internet em 2010	b_larga_fixa	0.961	0.194
A empresa utilizou banda larga móvel como conexão à internet em 2010	b_larga_movel	0.418	0.493
A empresa realizou compras de mercadorias ou serviços através da internet em 2010	compras	0.615	0.487
A empresa usou a internet para obter informações da Administração Pública em 2010	adm_publica	0.787	0.409
		16725	

Fonte: elaboração própria

4.2 Resultados

O objetivo desse trabalho foi aplicar alguns dos modelos de aprendizagem de máquina ao problema apresentado para compará-los e verificar qual a melhor alternativa no estudo da previsão de sinistros cibernéticos.

Com os resultados obtidos, a melhor proporção de concordância total, ou seja, a proporção de empresas no conjunto de teste que foram classificados como tendo um ataque (não ocorrência), realmente apresentando um ataque (não ocorrência), foi alcançada através do uso de redes neurais, atingindo 86%.

De acordo com os resultados apresentados na Tabela 3, observa-se que as redes neurais apresentam boa sensibilidade, ou seja, conseguem classificar com eficiência as empresas que declararam ter sofrido ataque. Tanto o valor de sensibilidade de 0,85 quanto a especificidade de 0,86 mostram que o modelo provou ser bastante eficiente na classificação de classes positivas e negativas.

O pior resultado, no entanto, foi com o uso do método GBM, com proporção de concordância total de 56%.

Tabela 3 – Resultado comparativo

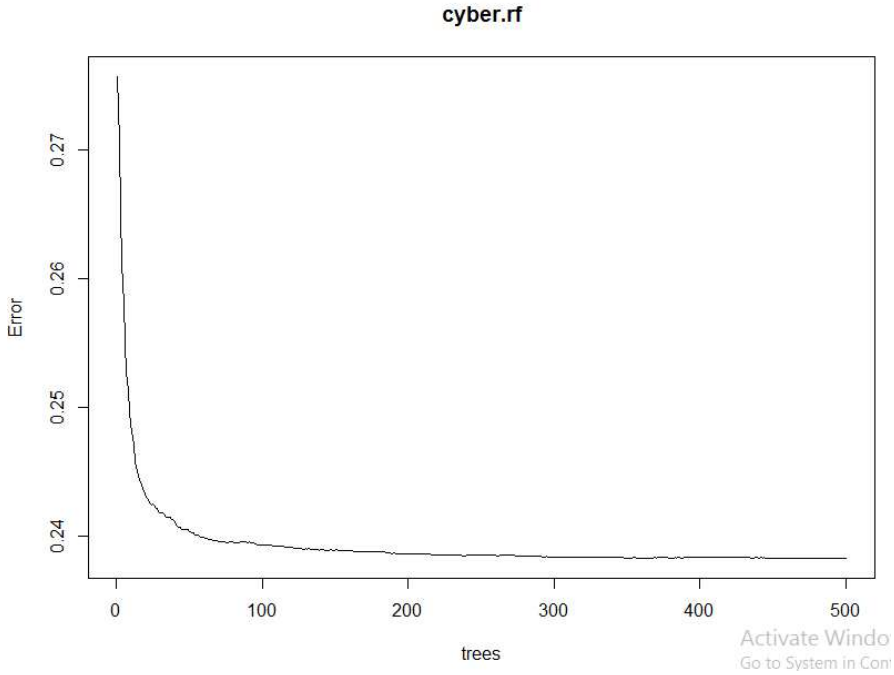
METODO	ACURACIA	ERRO	ESPECIFICIDADE	SENSIBILIDADE
Redes neurais	0,86	0,14	0,86	0,85
Árvore de Decisao	0,67	0,33	0,66	0,61
GAMLSS	0,61	0,39	0,94	0,10
GBM	0,56	0,44	0,42	0,68
Random forest	0,60	0,40	0,84	0,22

Fonte: elaboração própria

No que diz respeito a árvore de decisão e floresta aleatória, mesmo que seus desempenhos não tenham sido os melhores identificados, os métodos trazem, também, informações relevantes para o estudo. Ambos os métodos elencam a importância das variáveis na predição do sinistro cibernético. No caso de árvore de decisão, a variável que se mostrou mais importante foi "se empresa realizou compras de mercadorias ou serviços através da internet". Já na floresta aleatória, essa variável foi "se a empresa possuía departamento de TI".

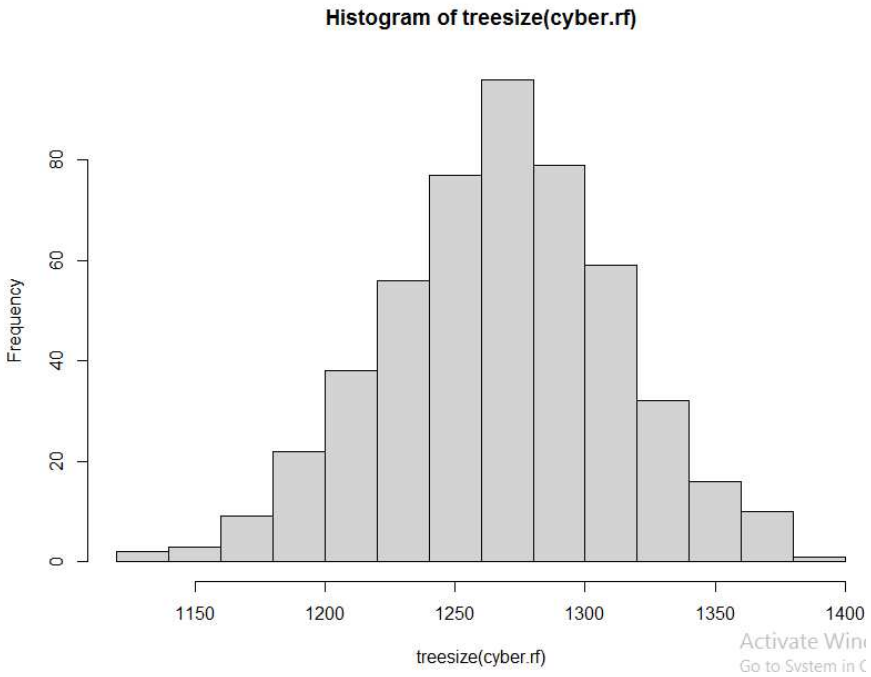
Ainda se tratando da floresta aleatória, o Gráfico 4 mostra como o erro evolui na medida em que se aumenta o número de árvores na simulação. O Gráfico 5 apresenta a distribuição do número de árvores utilizadas na simulação da floresta aleatória.

Gráfico 4 – Evolução do erro com o aumento do número de árvores



Fonte: elaboração própria

Gráfico 5 – Histograma da distribuição da quantidade de árvores



Fonte: elaboração própria

Em se tratando do método GAMLSS, método que também não apresentou bom resultado quando comparado às redes neurais, este traz também uma análise complementar ao ajustar os dados a uma distribuição de probabilidade, o que serve como suporte para decisões futuras em termos de precificação para o seguro cibernético. Para o caso em estudo, como visto na Tabela 4, a distribuição que melhor se ajustou aos dados de frequência de sinistros cibernéticos foi a distribuição Poisson Zero Inflada (ZIP). Essa distribuição tem dois parâmetros que se mostraram significantes no ajuste, como mostra a Tabela 5.

Tabela 4 – Ajuste de distribuição para frequência de sinistros

Distribuição	**Parâmetros**	**AIC**
ZIP	2	34693
ZAP	2	34693
ZANBI	3	34695
NBI	2	34823
PO	1	35334

Fonte: elaboração própria

Tabela 5 – Ajuste da distribuição ZIP

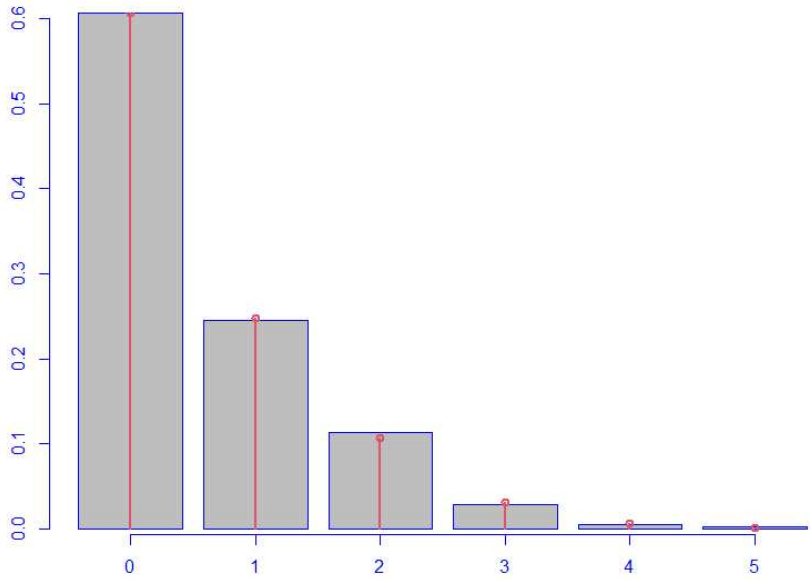
Parâmetro	Estimativa	Erro padrão	Valor t	Pr(>|t|)
μ	-0,1461473	0,0165688	-8,82061	< 2,22e-16 ***
σ	-0,7548475	0,0443263	-17,02934	< 2,22e-16 ***

Códigos de significância: 0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Fonte: elaboração própria

O Gráfico 6 mostra o ajuste gráfico dos dados de sinistros cibernéticos em relação à distribuição ZIP.

Gráfico 6 – Ajuste gráfico dos dados cibernéticos a distribuição ZIP



Fonte: elaboração própria

Sendo assim, os resultados deste trabalho estão de acordo com as afirmações de Meng (2011) de que modelos de aprendizagem de máquina são capazes de analisar padrões de tráfego de rede, identificar comportamentos suspeitos e até mesmo prever ataques antes que ocorram.

O fato é que o uso e estudo de aprendizagem de máquina pode ser bastante oportuno para auxiliar a resolver o problema apresentado, e muito possivelmente retirar o risco cibernético da lista dos dez principais riscos a serem combatidos globalmente – situação esta explicada por Georg (2023).

5 CONSIDERAÇÕES FINAIS

O estudo realizado destaca a significativa contribuição que os modelos de machine learning podem oferecer na mitigação de riscos cibernéticos. A superioridade das redes neurais na identificação proativa de possíveis ataques demonstra a eficácia desses métodos avançados na análise de dados complexos e na tomada de decisões preditivas, apesar dos resultados não tão promissores obtidos com uso do método GBM. No entanto, a observação do GAMLSS em ajustar os dados a distribuições de probabilidade, como a Poisson Zero Inflada (ZIP), oferece uma perspectiva adicional no desenvolvimento de estratégias de precificação de seguros cibernéticos mais embasadas.

Os resultados deste estudo indicam que as redes neurais são o melhor modelo de aprendizagem de máquina para a predição de sinistros cibernéticos. As redes neurais são capazes de aprender padrões complexos nos dados, o que lhes permite obter melhores resultados do que outros modelos.

No entanto, apesar do uso de uma larga amostra de dados, é importante ressaltar que os resultados deste estudo são baseados em um conjunto de dados específico. Para generalizar os resultados deste estudo, pode ser positivo realizar pesquisas adicionais com outros conjuntos de dados.

Além disso, é importante avaliar outros modelos de aprendizagem de máquina, como redes convolucionais e redes recorrentes. Esses modelos podem ser mais adequados para a predição de risco cibernético em certos contextos.

Para futuras pesquisas, recomenda-se a utilização de um conjunto de dados maior e mais diversificado. Além disso, recomenda-se a avaliação de outros modelos de aprendizagem de máquina, como redes convolucionais e redes recorrentes.

O entendimento dos limites e das capacidades desses modelos é fundamental para garantir não apenas a segurança digital, mas também para moldar estratégias empresariais robustas diante de um ambiente cada vez mais conectado e suscetível a ameaças cibernéticas em constante evolução.

REFERÊNCIAS

ABIODUN, Oludare Isaac et al. State-of-the-art in artificial neural network applications: A survey. **Heliyon**, v. 4, n. 11, 2018.

ANURADHA, C.; VELMURUGAN, T. A data mining based survey on student performance evaluation system. In: **2014 IEEE International Conference on Computational Intelligence and Computing Research**. IEEE, 2014. p. 1-4.

AYYADEVARA, V. Kishore; AYYADEVARA, V. Kishore. Gradient boosting machine. **Pro machine learning algorithms: A hands-on approach to implementing algorithms in python and R**, p. 117-134, 2018.

BENDOVSCHI, Andreea. Cyber-attacks–trends, patterns and security countermeasures. **Procedia Economics and Finance**, v. 28, p. 24-31, 2015.

BERNSTEIN, Peter L. **Desafio aos deuses: a fascinante história do risco**. Gulf Professional Publishing, 1997.

BIAU, Gérard; SCORNET, Erwan. A random forest guided tour. **Test**, v. 25, p. 197-227, 2016.

BÖHME, Rainer; LAUBE, Stefan; RIEK, Markus. A fundamental approach to cyber risk analysis. **Variance**, v. 12, n. 2, p. 161-185, 2019.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, p. 5-32, 2001.

CEBULA, James L.; YOUNG, Lisa R. **A taxonomy of operational cyber security risks**. Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst, 2010.

CHANG, L. I. N. Naive transfer learning approaches for suspicious event prediction. In: **2019 IEEE International Conference on Big Data (Big Data)**. IEEE, 2019. p. 5897-5901.

ELING, Martin; SCHNELL, Werner. What do we know about cyber risk and cyber risk insurance?. **The Journal of Risk Finance**, v. 17, n. 5, p. 474-491, 2016.

GEORG, Marcus Aurélio Carvalho. **Proposta de Modelo de Mensuração de Apetite a Riscos Cibernéticos: Uso do Método AHP e da Estrutura Básica de Segurança Cibernética.** 2023. Dissertação de Mestrado.

GIGER, Maryellen L. Machine learning in medical imaging. **Journal of the American College of Radiology**, v. 15, n. 3, p. 512-520, 2018.

GRĄBCZEWSKI, Krzysztof; GRĄBCZEWSKI, Krzysztof. Unified View of Decision Tree Induction Algorithms. **Meta-Learning in Decision Tree Induction**, p. 119-137, 2014.

HANDA, Anand; SHARMA, Ashu; SHUKLA, Sandeep K. Machine learning in cybersecurity: A review. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 9, n. 4, p. e1306, 2019.

JUSTO, Mariana Martins da Cruz. **Risco cibernético e Regulamento Geral de Protecão de Dados, adaptação das empresas à nova realidade**. 2018. Tese de Doutorado.

KUMAR, Deepak; SARANGI, Pradeepta Kumar; VERMA, Rajit. A systematic review of stock market prediction using machine learning and statistical techniques. **Materials Today: Proceedings**, v. 49, p. 3187-3191, 2022.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **nature**, v. 521, n. 7553, p. 436-444, 2015.

MAHESH, B. Machine Learning Algorithms - A Review. **International Journal of Science and Research (IJSR)**, v. 9, n. 1, p. 381-386, 2020.

MARLETTA, Andrea. ROC Curve in GAMLSS as Prediction Tool for Big Data. In: **Data Science and Social Research II: Methods, Technologies and Applications**. Springer International Publishing, 2021. p. 247-257.

MAROTTA, Angelica et al. Cyber-insurance survey. **Computer Science Review**, v. 24, p. 35-61, 2017.

MENG, Yu-Xin. The practice on using machine learning for network anomaly intrusion detection. In: **2011 International Conference on Machine Learning and Cybernetics**. IEEE, 2011. p. 576-581.

MIENYE, Ibomoiye Domor; SUN, Yanxia; WANG, Zenghui. Prediction performance of improved decision tree-based algorithms: a review. **Procedia Manufacturing**, v. 35, p. 698-703, 2019.

RAJKOMAR, Alvin; DEAN, Jeffrey; KOHANE, Isaac. Machine learning in medicine. **New England Journal of Medicine**, v. 380, n. 14, p. 1347-1358, 2019.

ROMANOSKY, Sasha et al. Content analysis of cyber insurance policies: How do carriers price cyber risk?. **Journal of Cybersecurity**, v. 5, n. 1, p. tyz002, 2019.

SHINDE, P.; SHAH, S. A Review of Machine Learning and Deep Learning Applications. **Institute of Electrical and Electronics Engineers**, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018.

SPEISER, Jaime Lynn et al. A comparison of random forest variable selection methods for classification prediction modeling. **Expert systems with applications**, v. 134, p. 93-101, 2019.

UMA, M.; PADMAVATHI, Ganapathi. A survey on various cyber attacks and their classification. **Int. J. Netw. Secur**., v. 15, n. 5, p. 390-396, 2013.

WHIG, Pawan. Artificial intelligence and machine learning in business. **International Journal on Integrated Education**, v. 2, n. 2, p. 334128, 2021.

WU, Yu-chen; FENG, Jun-wen. Development and application of artificial neural network. **Wireless Personal Communications**, v. 102, p. 1645-1656, 2018.

YADAV, Neha et al. **An introduction to neural network methods for differential equations**. Berlin: Springer, 2015.

ZHOU, Zhi-Hua. **Machine learning**. Springer Nature, 2021.