UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS RUSSAS

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CÍCERO ROMÃO RIBEIRO PEREIRA FILHO

SURVEY: VIABILIDADE E IMPACTO NA IMPLANTAÇÃO DE UM MODELO DE PREDIÇÃO DE ATAQUES DE REDES DE COMPUTADORES

RUSSAS-CE

2023

CÍCERO ROMÃO RIBEIRO PEREIRA FILHO

SURVEY: VIABILIDADE E IMPACTO NA IMPLANTAÇÃO DE UM MODELO DE PREDIÇÃO DE ATAQUES DE REDES DE COMPUTADORES

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientador: Prof. Dr. Reuber Regis De Melo

RUSSAS-CE

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)

P49s Pereira Filho, Cícero Romão Ribeiro.
Survey : viabilidade e impacto na implantação de um modelo de predição de ataques de redes de computadores / Cícero Romão Ribeiro Pereira Filho. – 2023.
34 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Russas, Curso de Ciência da Computação, Russas, 2023.
Orientação: Prof. Dr. Reuber Regis De Melo.

1. survey. 2. cybersegurança. 3. redes de computadores. I. Título.

CDD 005

CÍCERO ROMÃO RIBEIRO PEREIRA FILHO

SURVEY: VIABILIDADE E IMPACTO NA IMPLANTAÇÃO DE UM MODELO DE PREDIÇÃO DE ATAQUES DE REDES DE COMPUTADORES

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Reuber Regis De Melo (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Pablo Luiz Braga Soares
Universidade Federal do Ceará - UFC

Prof. Dr. Bonfim Amaro Júnior
Universidade Estadual do Ceará - UECE

À minha família, por sua capacidade de acreditar em mim e investir em mim. Mãe, seu cuidado e dedicação foi que deram, em alguns momentos, a esperança para seguir. Pai, sua presença significou segurança e certeza de que não estou sozinho nessa caminhada.

“ Conhecimento dá poder, mas só o caráter grangeia respeito. ”

(Bruce Lee)

RESUMO

A tecnologia da informação está enraizada em uma grande parte da sociedade atuando em diferentes áreas e apresentando uma evolução constante, a qual se deve principalmente pela rede de computadores que possibilita que os usuários e máquinas se conectem. A rede é de vital importância na atualidade sendo um meio de obter, compartilhar e armazenar informações essas que podem possuir um valor imensurável. Perante isso é perceptível que indivíduos mal-intencionados podem almejar essas informações esses indivíduos em sua maioria possuem conhecimentos para invadir redes de computadores e obter esses dados, os mesmo também são conhecidos como *hackers*. Em razão desses ataques técnicas para garantir a segurança foram desenvolvidas com o tempo, em sua maioria esses métodos são corretivos sendo aplicados durante ou depois dos episódios de ataques cibernéticos. Desse modo o objetivo é apresentar a pesquisa para viabilização da implantação do modelo de predição de ataques de rede e o impacto que o modelo poderá causar. O modelo de pesquisa utilizado foi o *survey* realizando o processo de seleção de amostra utilizando técnicas para separar a população alvo de maneira efetiva e concisa, para que assim os dados coletados sejam válidos. À vista disto o questionário utilizado foi construído objetivando os principais pontos que necessitam de validação como viabilidade, impacto, custo e catalogação dos indivíduos da amostra. Por fim alguns dados estatísticos foram obtidos como resultado do *survey* denotando uma breve discussão sobre as peculiaridades.

Palavras-chave: survey; segurança de redes; ataques cibernéticos; previsão de ataques; aprendizado de máquina; redes neurais;

ABSTRACT

Information technology is rooted in a large part of society, working in different areas and showing constant evolution, which is mainly due to the computer network that allows users and machines to connect. The network is of vital importance today as a means of obtaining, sharing and storing information that can be of immeasurable value. In view of this, it is noticeable that malicious individuals can target this information, these individuals mostly have the knowledge to invade computer networks and obtain this data, they are also known as *hackers*. Because these attack techniques to ensure security have been developed over time, most of these methods are correctives being applied during or after episodes of cyber attacks. Thus, the objective is to present the research to enable the implementation of the network attack prediction model and the impact that the model may cause. The research model used was the *survey* performing the sample selection process using techniques to separate the target population in an effective and concise way, so that the collected data are valid. In view of this, the questionnaire used was constructed aiming at the main points that need validation, such as feasibility, impact, cost and cataloging of the individuals in the sample. Finally some statistical data were obtained as a result of the *survey* denoting a brief discussion about the peculiarities.

Keywords: survey; network security; cyber attacks; attack prediction; apprenticeship of machine; neural networks;

LISTA DE ILUSTRAÇÕES

LISTA DE SÍMBOLOS

HTTPS	Hyper Text Transfer Protocol Secure
IP	Internet Protocol
CSA	Cyber Situational Awareness
ELK	Elasticsearch, Logstash e Kibana.
UDP	User Datagram Protocol
TCP	Transmission Control Protocol/Internet Protocol
SVM	Support Vector Machines
RNA	Rede Neural Artificial
HIDS	The Host-Based
NIDS	The network-based
API	application program interface

SUMÁRIO

1 INTRODUÇÃO

As redes de computadores têm crescido rapidamente, e com isso permitiu que várias aplicações importantes fossem criadas na área empresarial, artística, educacional, governamental e científica. Nesse avanço teve o crescimento contínuo da internet ligando os diversos tipos de redes de computadores em escala global, levando comunicação e informação a milhares de pessoas COMER2016. A informação se tornou o bem mais importante da atualidade.

Nesse contexto a informação se tornou alvo de indivíduos maliciosos que pretendem obtê-la de maneira ilícita, utilizando de seus conhecimentos sobre a tecnologia da informação para desenvolver técnicas objetivando encontrar brechas nas redes invadindo-as e adquirindo os dados almejados.

De acordo com Lyon (2008) dentre os meios utilizados pelos *hackers* (indivíduos com o intuito de invadir redes) estão ataques de varredura os quais pretendem escanear as portas de uma rede fornecendo assim informações sobres quais programas estão rodando em cada porta para assim os hackers aproveitarem de suas vulnerabilidades, ataques de força bruta como citado em Knudsen e Robshaw (2011) que se utilizam de uma *wordlist*(conjunto de caracteres e palavras-chave) para gerar possíveis senhas até encontrar a senha do seu alvo.

Em vista disto o perigo constante que entorna o contexto das redes de comunicação impulsionou o surgimento de técnicas para garantir a eficiência na proteção de informações. Destacam-se *softwares* que escaneiam as máquinas para detectar possíveis programas maliciosos ou vírus relatando assim ao usuário dos perigos e tratando de alguns deles. Além de informar sobre procedimentos utilizados para manter a segurança da rede e dos dispositivos, por meio do bloqueio da instalação de aplicativos sem a análise de um especialista, assim como o bloqueio de sites que possam ser maliciosos. Porém, todas as técnicas citadas anteriormente assim como similares, buscam prevenir ou solucionar problemas já existentes.

Diante disto, esse trabalho visa investigar através da metodologia de pesquisa *survey* a viabilidade e o impacto do modelo de predição de ataques de rede utilizando netflow. O modelo permite ao usuário a proteção contra possíveis invasões por meio da predição do fluxo de ataques a rede. A pesquisa almeja coletar e analisar os dados buscando a validação para aplicação do modelo, assim como o impacto causado pelo mesmo na indústria. Por meio da metodologia de pesquisa *survey*, aplicando um questionário que objetiva validar as informações referentes ao modelo por meio da coleta de respostas direcionadas a uma amostra estruturada para a pesquisa.

1.1 Objetivos

O objetivo principal desse trabalho é apresentar um *survey* baseado no modelo de predição de ataques de rede utilizando *netflow*, e denotar gráfica e estatisticamente os resultados obtidos seguidos pela conclusão alcançada.

Os objetivos secundários os quais serão fundamentais para alcançar o objetivo principal são:

- Apresentar a base teórica necessária para compreensão do funcionamento e aplicação do modelo de predição.
- Definir o público-alvo, para que o *survey* seja aplicado na amostra deste público.
- Criar um questionário com itens que possam levantar pontos cruciais para a análise de viabilidade e impacto da pesquisa.
- Analisar e apresentar os dados coletados por meio da pesquisa utilizando o questionário.
- Apresentar a metodologia utilizada na pesquisa permitindo assim a replicação do processo.

1.2 Estrutura do Trabalho

A estrutura do trabalho composta por este capítulo inicial onde temos a motivação para o desenvolvimento assim como os objetivos almejados.

Em seguida no capítulo 2, é introduzida toda base teórica essencial para o entendimento e desenvolvimento do trabalho. São apresentadas as categorias de ataques de rede assim como suas características. Também são apresentadas as ferramentas para gerenciamento e monitoramento de rede e suas funcionalidades. Por fim no capítulo é apresentado a fundamentação de redes neurais, aprendizado profundo e modelos de predição que serão utilizados no trabalho.

Em seguida no capítulo 3 é apresentada toda estrutura metodológica utilizada no trabalho, desde o funcionamento e arquitetura do survey para pesquisa e coleta de dados, até a estrutura do modelo de predição e as características de seus componentes.

Conseguinte são apresentados no capítulo 4 os resultados obtidos, assim como uma breve análise.

Por fim no capítulo 5 é apresentada a conclusão resultante da pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a base teórica e conceitos fundamentais para a compreensão desse trabalho. A seção 2.1 traz os conceitos básicos sobre o funcionamento e importância da pesquisa de levantamento do tipo *survey*. A seção 2.2 apresenta o conceito de amostragem, desde a seleção da população alvo que garante a validade dos resultados coletados até as formas de seleção de amostra. A seção 2.3 apresenta as duas vertentes de modelos de pesquisa assim como suas características e aplicações. Complementando a seção 2.4 denota as diversas maneiras utilizadas para coleta de dados nas pesquisas de levantamento, assim como os principais pontos a serem considerados na estruturação das perguntas. Conseguinte a seção 2.5 apresenta as tecnologias e conceitos utilizados no modelo de estudo, assim como o funcionamento e o papel de cada uma dessas tecnologias. Por fim a seção 2.6 apresenta modelos que podem ser utilizados para a predição de ataques, os conhecimentos básicos sobres Redes Neurais, e conceitos básicos sobre aprendizado profundo.

2.1 Metodologia de Pesquisa *Survey*

A pesquisa de levantamento é um tipo específico de estudo de campo que envolve a coleta de dados de uma amostra de elementos extraídos de uma população bem definida por meio do uso de um questionário. A pesquisa pode ser valiosa para os psicólogos sociais e, útil para diferentes objetivos de estudos como apresentado em (VISSER *et al.*, 2000). As informações são coletadas por meio de procedimentos padronizados para que cada indivíduo seja indagado sobre as mesmas perguntas e de maneira semelhante. O objetivo do *survey* é coletar respostas de indivíduos distintos que fazem parte de uma mesma amostra, assim obtendo um resultado válido para a população assim como foi dissertado em (SCHEUREN, 2004).

A pesquisa *survey* oferece uma maneira eficiente e sistemática de coletar dados quantitativos em grande escala. Porém é importante considerar suas limitações, como possíveis vieses de amostragem, questões de validade e confiabilidade das respostas e dificuldades na formulação de perguntas. Essas considerações devem ser cuidadosamente abordadas durante o planejamento e a implementação da pesquisa Freitas *et al.* (2000). Logo a estrutura do *survey* segue os seguintes passos:

- Definição dos objetivos da pesquisa: Estabelecer claramente os objetivos da pesquisa, identificando quais questões precisam ser respondidas e que tipo de informações são

necessárias.

- Desenvolvimento do questionário: O questionário deve ser elaborado, com perguntas claras e objetivas. Pode conter perguntas abertas ou perguntas fechadas. O questionário deve ser testado antes da implementação para garantir sua clareza e eficácia.
- Seleção da amostra: É necessário determinar o tamanho e o método de seleção da amostra. A amostra deve ser representativa da população-alvo, garantindo que todos os grupos relevantes estejam adequadamente representados.
- Coleta de dados: Os dados podem ser coletados por meio de entrevistas presenciais, entrevistas por telefone, questionários online ou por correio. A coleta de dados pode ser realizada por entrevistadores treinados ou de forma autodirigida, dependendo do método escolhido.
- Análise dos dados: Após a coleta de dados, é realizada a análise estatística para extrair informações relevantes. Isso pode envolver técnicas como tabulação de dados, cálculo de frequências, análise de correlação e testes estatísticos, dependendo das questões de pesquisa e das variáveis envolvidas.
- Relatório dos resultados: Os resultados da pesquisa são compilados e apresentados em um relatório, que pode incluir tabelas, gráficos e análises descritivas. O relatório deve ser claro, objetivo e acessível aos leitores, permitindo a interpretação e a tomada de decisões com base nos resultados apresentados.

2.2 Amostragem

O levantamento é utilizado quando se quer saber de que maneira determinados comportamentos aparecem em certo conjunto de pessoas para o qual se vai generalizar essa descoberta. Ao conjunto de todas as pessoas que têm ao menos uma característica em comum dá-se o nome de população. O motivo pelo qual é feita a utilização de uma amostra em vez de toda a população é a exigência de tempo e de redução de custo.

A amostra deve realmente representar a população em estudo, e os participantes voluntários só devem ser aceitos quando estiverem nessa condição. O ideal é que sejam utilizadas formas que não enviesem a amostra. Embora se pesquise uma amostra da população, o resultado não é menos preciso do que se fosse investigar toda a população. Nesse método de pesquisa, não se inferem as causas da presença ou ausência de determinado comportamento na amostra. As variáveis se apresentam de forma natural, e o pesquisador não pode manuseá-las como em pes-

quisas de laboratório. As respostas dadas por aquela determinada amostra são estendidas para a população por meio de estudos estatísticos probabilísticos. Assim, declara-se que provavelmente, a partir daquela amostra pesquisada, toda a população da qual aquela amostra faz parte apresenta a mesma resposta (BAPTISTA, 2016).

Existem diferentes métodos de amostragem, cada um com suas características e aplicações específicas Bolfarine e Bussab (2005). Aqui estão alguns dos principais tipos de amostragem:

- Amostragem Aleatória Simples: Nesse método, cada elemento da população tem a mesma probabilidade de ser selecionado para fazer parte da amostra. É um dos métodos mais básicos e amplamente utilizados. Pode ser realizado por meio de sorteio aleatório, usando números aleatórios ou uma tabela de números aleatórios.
- Amostragem Estratificada: Nesse método, a população é dividida em grupos ou estratos com características semelhantes. Em seguida, uma amostra é selecionada aleatoriamente de cada estrato, proporcional ao tamanho ou importância de cada grupo. A amostragem estratificada garante que cada grupo seja representado adequadamente na amostra final, o que pode ser útil quando existem diferenças significativas entre os estratos.
- Amostragem por Conglomerados: Nesse método, a população é dividida em grupos maiores chamados conglomerados. Em vez de selecionar indivíduos individuais, a amostra consiste em selecionar aleatoriamente alguns conglomerados e, em seguida, coletar dados de todos os indivíduos dentro desses conglomerados selecionados. Esse método é eficiente quando a população é grande e dispersa geograficamente.
- Amostragem Sistemática: Nesse método, os elementos da população são ordenados de alguma forma (por exemplo, por número de identificação) e, em seguida, é selecionado um elemento inicial aleatório. A partir desse ponto de partida, os elementos são selecionados de forma sistemática, com intervalos regulares, até que a amostra desejada seja alcançada. Esse método é útil quando a população está organizada em uma ordem específica.
- Amostragem por Conveniência: Esse método envolve a seleção dos elementos mais acessíveis ou convenientes para a amostra. Embora seja rápido e fácil de implementar, pode levar a um viés na seleção, pois os elementos selecionados podem não ser representativos da população em estudo.

A escolha do método de amostragem depende do objetivo do estudo, da disponibilidade de recursos e da natureza da população. Além disso, é importante garantir que a amostra

seja suficientemente grande e representativa para que os resultados sejam aplicáveis para a população maior.

2.3 Pesquisa quantitativa e qualitativa

Na pesquisa científica, encontram-se dados que podem ser quantificados e dados que podem ser analisados de forma qualitativa. A pesquisa quantitativa é realizada por meio da coleta de dados utilizados para medir variáveis, esse modelo é utilizado para criar uma base e obter conclusões gerais sobre o objeto de pesquisa visto que os resultados obtidos são conclusivos e estatísticos. O modelo de pesquisa qualitativo busca coletar informações que descrevem a experiência relacionada ao tema ao invés de uma medição, esse tipo de pesquisa usa gráficos ou tabelas para medir opiniões, pontos de vista e atributos de forma numérica. Os métodos de pesquisa qualitativa envolvem observação direta, como entrevistas e grupos focais e pesquisa de mercado (WAINER *et al.*, 2007).

2.4 Técnicas de coleta de dados

As principais técnicas de coletas de dados são questionários por telefone, pessoalmente, por e-mail, pelo correio, entrevistas estruturadas, semiestruturadas, não estruturadas, face a face, por telefone, gravadas. Ao se montar um questionário, deve-se atentar, em primeiro lugar, para a clareza das questões para que o participante tenha facilidade de responder. Muitas vezes, utilizam-se juízes (colegas pesquisadores, colegas de profissão) para atestarem se houve exata compreensão do que foi perguntado; esse procedimento denomina-se validação de conteúdo com precisão de juízes. Eles irão verificar a adequação do vocabulário empregado, a precisão dos enunciados, a pertinência do material em relação ao domínio previamente definido e a possibilidade de vieses (BAPTISTA, 2016).

2.5 Ferramentas de gestão de ataques a redes

O *Elastic Stack*, outrossim conhecido como ELK, cujo significado vem de ser um acrônimo para três projetos emphopen source: *Elasticsearch*, emphLogstash e *Kibana*. O *Elasticsearch* é um mecanismo de busca e análise Gupta e Gupta (2017). *Logstash* é um *pipeline* de processamento de dados do lado do servidor que faz a ingestão de dados a partir de inúmeras fontes simultaneamente, transforma-os e envia-os para um "esconderijo"como o

Elasticsearch. Kibana permite que os usuários visualizem dados utilizando diagramas e gráficos no *Elasticsearch* (AGGARWAL, 2022a).

Conhecido por suas REST APIs simples, natureza distribuída, velocidade e escalabilidade, o Elasticsearch é o componente central do Elastic Stack, um conjunto de ferramentas *opensource* para ingestão, enriquecimento, armazenamento, análise e visualização de dados (KUC; ROGOZINSKI, 2013).

Por sua vez o *Beats* a plataforma *opensource* para agentes de dados de finalidade única. Eles enviam dados de centenas ou milhares de computadores e sistemas para o Logstash ou o Elasticsearch como em (AGGARWAL, 2022b). Assim como os diferentes tipos de dados temos também as derivações do *Beats*, como: auditbeat, filebeat, functionbeat, heartbeat, metricbeat, packetbeat, winlogbeat.

Merece destaque também, o kibana, um aplicativo que permite visualizar os dados armazenados no *Elastic Search*. A ferramenta permite que você escolha como seus dados serão apresentados de forma visual. Podendo assim serem representados por diversas formas como, por exemplo: histograma, gráfico de linhas, setor, gráfico de dispersão, etc. Também é possível definir exibições personalizadas. Por fim, o Kibana é interativo para fornecer aos usuários os meios para obter um alto nível de percepção dos dados que estão sendo analisados (GUPTA, 2015).

A tecnologia chamada *Netflow* que foi criada pela empresa Cisco se faz presente em alguns roteadores, essa mesma tecnologia permite que esses roteadores possuam a habilidade de capturar dados de pacotes de entrada e saída, esses dados capturados chamados fluxos são armazenados em *cache*. Quando o fluxo principal se encerra os dados capturados são exportados para um coletor também conhecido como dispositivo de coleta. Os dados capturados são exportados pelo Netflow no formato de datagramas UDP, os datagramas podem ser compatíveis com as seguintes versões do Netflow: versão 1, versão 5, versão 7, versão 8 ou versão 9 (SYSTEMS, 2016).

2.6 Modelos de predição e aprendizagem de máquina

A seção apresenta a base teórica necessária sobre os modelos de predição de ataques, seus funcionamentos e peculiaridades. Assim como a estrutura e atividade das redes neurais e aprendizado de máquina.

2.6.1 Predição Baseada em Modelos Discretos

Em um sistema de eventos discretos, um ou mais fenômenos de interesse mudam seu valor, ou estado, em pontos discretos (ao invés de continuamente) no tempo. A ocorrência destes eventos muda o estado do sistema em cada momento. Dessa forma, assumimos não haver mudanças no sistema entre um evento e outro. Mesmo em caso de haver incrementos fixos de avanço no tempo, o que não é muito comum, a evolução do sistema não ocorre de forma contínua no tempo. Na simulação por eventos discretos, o tempo é dividido em pequenas fatias e o estado do sistema é atualizado conforme as atividades que ocorrem em cada fatia do tempo. Como nem toda fatia de tempo possui ocorrência de atividade, esta simulação é mais rápida que a simulação contínua. A técnica do próximo evento possui duas vantagens sobre a técnica de fatiamento de tempo. A primeira é que o incremento do tempo é ajustado automaticamente há períodos com alto ou baixo índice de ocorrência de atividades evitando desperdício, ou verificações desnecessárias do estado do modelo.

Conseguinte a segunda é que a abordagem do próximo-evento deixa claro onde ocorre aproximadamente eventos. Ela é geral e engloba a técnica do fatiamento de tempo. O contrário é falso. Ao invés de usar somente o paradigma de programação estruturada baseada em eventos, a simulação por eventos discretos pode ser baseada em eventos, atividades ou em processos. Os elementos-base de uma simulação por eventos discretos são: o estado, o relógio e a lista de eventos. O estado do evento é representado por variáveis que representam as propriedades do sistema a ser estudado. O relógio mantém o controle da evolução temporal da simulação na unidade de tempo escolhida. A lista de eventos é chamada lista de eventos pendentes no início da simulação.

À medida que o relógio da simulação avança, os eventos são realizados e o estado do sistema é atualizado. Os eventos pendentes são organizados em uma lista de prioridades, ordenados por duração do evento. Independentemente de como são ordenados, os eventos são removidos da lista na ordem cronológica da simulação. A partir deste momento, a simulação computa as estatísticas do sistema, que quantifica os aspectos de interesse. Em um modelo de simulação, as estatísticas não são derivadas de distribuições de probabilidade, mas de médias de replicações de rodadas do modelo. Para avaliar a qualidade do resultado, intervalos de confiança são construídos. O fim da simulação pode ser determinado por tempo de simulação ou por uma medida estatística (FLáVIO, 2022).

2.6.2 Predição Baseada em Modelos Contínuos

Dentro desta categoria se enquadram os métodos baseados em séries temporais e modelos cinza. Séries temporais representam uma ferramenta muito interessante para análise preditiva, utilizadas em vários campos, incluindo segurança cibernética. São também frequentemente empregadas na detecção de anomalias. Uma série temporal representa padrões de tráfego de rede comum. Posteriormente, os desvios que não coincidem com os valores esperados de tráfego de rede em um determinado momento são proclamados como uma anomalia. Embora a terminologia e os métodos de detecção de anomalias sejam semelhantes à previsão de ataque, os dois casos de uso são substancialmente diferentes. Os modelos cinza são normalmente usados para prever situações de segurança cibernética e definem outro exemplo de metodologias que empregam um modelo matemático contínuo. Na terminologia da teoria de cinza, uma situação sem informação é definida como preta e uma situação com informações completas como branca. Como às duas opções são idealizadas, os problemas reais estão em algum lugar no meio, em uma situação definida como cinza (CALIS, 2021).

2.6.3 Redes Neurais

Uma rede neural artificial (RNA) tem duas facetas elementares: a arquitetura e o algoritmo de aprendizagem. Essa divisão surge naturalmente pelo paradigma como a rede é treinada. Ao contrário de um computador com arquitetura de von Neumann que é programado, a rede é treinada por exemplos de treino. O conhecimento sobre o problema em consideração está guardado nos exemplos que têm que estar obrigatoriamente disponíveis. O algoritmo de aprendizagem generaliza esses dados e memoriza o conhecimento nos parâmetros adaptáveis da rede, os pesos. Assim o construtor de um sistema baseado em RNA tem dois graus de liberdade, a definição sobre a categoria de rede para resolver o problema em consideração e o algoritmo para treinar a rede, i.e. para adaptar os pesos da rede. A composição da rede é feita pelos neurônios. Normalmente o tipo de processamento de um único neurônio é a combinação linear das entradas com os pesos seguida pela passagem da combinação linear por uma função de ativação. O problema a ser resolvido normalmente define restrições em relação aos tipos de redes e algoritmos de aprendizagem possíveis. Neste texto distinguem-se redes com propagação do fluxo de informação para frente, redes recorrentes (com realimentação das saídas para as entradas) e redes competitivas. Em relação aos algoritmos de adaptação, vamos distinguir entre

aprendizagem supervisionada e aprendizagem não-supervisionada (RAUBER, 2005).

2.6.4 Aprendizado Profundo

O aprendizado profundo é o subcampo da inteligência artificial que se concentra na criação de grandes modelos de redes neurais capazes de tomar decisões precisas baseadas em dados. O aprendizado profundo é particularmente adequado para contextos em que os dados são complexos e onde há grandes conjuntos de dados disponíveis Kelleher (2019). No aprendizado profundo são utilizados sistemas com múltiplas camadas podem realizar de forma integrada a extração das características mais relevantes e detecção de objetos. Os algoritmos de Aprendizado Profundo utilizam como base este conceito, diferenciando-se entre si pela estruturação da arquitetura, profundidade da arquitetura, os tipos de camada usados e os parâmetros de inicialização de cada modelo Oliveira *et al.* (2018).

O aprendizado profundo permite que modelos computacionais compostos de várias camadas de processamento aprendam representações de dados com vários níveis de abstração. Esses métodos melhoraram drasticamente o estado da arte em reconhecimento de fala, reconhecimento visual de objetos, detecção de objetos e muitos outros domínios, como descoberta de drogas e genômica. O aprendizado profundo descobre uma estrutura complexa em grandes conjuntos de dados usando o algoritmo de retro propagação para indicar como uma máquina deve alterar seus parâmetros internos usados para calcular a representação em cada camada a partir da representação na camada anterior. Redes convolucionais profundas trouxeram avanços no processamento de imagens, vídeo, fala e áudio, enquanto as redes recorrentes iluminaram dados sequenciais, como texto e fala (LECUN *et al.*, 2015). Esses algoritmos são projetados para aprender e extrair representações complexas dos dados, permitindo o processamento de informações de forma hierárquica e não linear.

Redes Neurais Convolucionais (CNNs) são amplamente utilizadas para tarefas de processamento de imagens e visão computacional. Elas são projetadas para reconhecer padrões espaciais em dados, como pixels de uma imagem, e são compostas por camadas convolucionais, de *pooling* e completamente conectadas. CNNs têm sido aplicadas com sucesso em tarefas como classificação de imagens, detecção de objetos e reconhecimento facial Juraszek *et al.* (2014).

Redes Neurais Recorrentes (RNNs) são adequadas para processar dados sequenciais, como séries temporais ou texto. A principal característica das RNNs é a capacidade de manter informações de estado oculto, permitindo que informações anteriores influenciem a saída atual.

Essa propriedade é útil em tarefas como tradução automática, reconhecimento de voz e geração de texto Carvalho *et al.* (2018).

Redes Neurais Generativas Adversariais (GANs) são compostas por duas redes neurais, o gerador e o discriminador, que competem entre si. O gerador gera amostras sintéticas, enquanto o discriminador tenta distinguir entre amostras reais e falsas. Esse jogo entre o gerador e o discriminador resulta em um processo de treinamento no qual o gerador aprende a gerar amostras cada vez mais realistas. GANs são amplamente utilizadas em tarefas de geração de imagens, como criação de rostos sintéticos e aprimoramento de imagens Zuba *et al.* (2021).

Redes Neurais Autoencoders são redes neurais projetadas para aprender representações compactas e eficientes de dados de entrada. Eles consistem em duas partes principais: um codificador que mapeia os dados para um espaço de menor dimensionalidade e um decodificador que tenta reconstruir os dados originais a partir dessa representação. Autoencoders têm aplicações em compressão de dados, remoção de ruído e reconstrução de imagens Gilbert *et al.* (2023).

3 METODOLOGIA

Neste capítulo, é apresentada a metodologia para realização do *survey*. Desde a definição do objeto de pesquisa, instrumento para coleta de dados e seleção da amostra. Almejando por fim apresentar o escopo completo da pesquisa viabilizando assim a replicabilidade do modelo de pesquisa.

3.1 Definição dos Objetivos da Pesquisa

As principais questões a serem respondidas pelo *survey*, estão relacionadas a viabilidade, impacto e custos do modelo de predição de ataques de rede. Foi realizada uma avaliação da possibilidade de implementação e do impacto do modelo de previsão de ataques de rede, nessa análise foram consideradas as possíveis mudanças ocasionadas pela implementação do modelo, a dificuldade de implementação, o nível de conhecimento dos participantes da amostra alvo sobre as áreas correlatas ao modelo e o custo para implantação.

Também foi considerado nesta pesquisa o impacto do modelo de predição de ataques de rede na detecção e prevenção de ataques cibernéticos. Foram explorados os benefícios que esse modelo oferece, como a melhoria da segurança cibernética e a redução dos riscos associados aos ataques de rede.

3.2 Desenvolvimento do Instrumento de Coleta de Dados

Para entender se o modelo é viável, seu impacto na detecção/prevenção de ataques e os custos envolvidos. Foi utilizado um questionário estruturado, as respostas serão tratadas com confidencialidade. Para aplicação foi utilizado a via digital por meio da distribuição google forms, possibilitando a escalabilidade de crescimento da amostra. A seguir será apresentado o questionário utilizado no *survey*.

Através do uso da distribuição google forms foi possível a coleta de resultados representados de forma gráfica, facilitando a análise e representatividade dos resultados obtidos, em vista que para a análise estatística os gráficos possibilitaram uma visão bem definida dos dados Cícero (2023).

3.3 Seleção da Amostra

A amostragem é importante na pesquisa, pois garante a representatividade dos participantes e a diversidade dos resultados. Uma amostra apropriada foi cuidadosamente selecionada para estudar a viabilidade e o impacto do modelo de previsão de ataques cibernéticos.

Uma combinação de técnicas de amostragem foi usada para construir a amostra. Primeiro foi definido o tipo de pesquisa apropriada para validar o objeto de pesquisa, visando obter a viabilidade de implantação e possível impacto causado para um modelo apenas teórico, foi selecionado o tipo de pesquisa quantitativa. Em seguida foi utilizada a amostragem aleatória para garantir que cada indivíduo na população-alvo tenha uma chance igual de ser selecionado. Essa abordagem ajuda a garantir a representatividade dos participantes. Também foi usado a amostragem estratificada para dividir a população em diferentes grupos com base em características relevantes para o objeto de pesquisa. Os grupos representados são:

- Estudante
- Profissional na área de tecnologia
- Profissional em outra área
- Outros

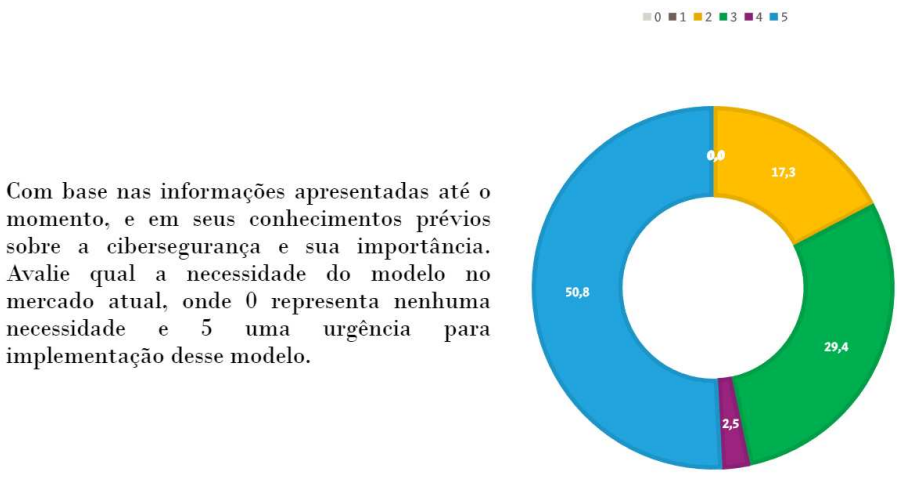
4 RESULTADOS

Neste capítulo, serão apresentados os resultados obtidos através do *survey*. Ao longo desta pesquisa foi mantido como foco principal obter dados mediante uma análise quantitativa para apoiar positivamente os pontos focais deste *survey* que foram a viabilidade do projeto e os possíveis impactos causados pelo modelo. A seguir, serão apresentados em destaque os principais resultados obtidos pela pesquisa, discutindo suas implicações e relevância para o objetivo deste *survey*. Vale ressaltar que foram coletadas um total de 197 respostas para o questionário totalizando assim o tamanho da amostra.

4.1 Viabilidade de implementação do modelo

Nesta seção, serão apresentados os principais resultados obtidos referentes a viabilidade de implementação do modelo de predição de ataques de rede. Entre as questões presentes no questionário utilizado para coleta de dados da pesquisa, a 2, 6 e 10 são relacionadas diretamente a viabilidade. De início podemos considerar a questão de número 2, ilustrado na Figura 1, que avalia a necessidade de implementação do modelo no mercado atual onde foi obtido um retorno positivo, em vista que a questão utiliza uma classificação de escala de 0 a 5 onde quanto maior o número maior a necessidade. Das respostas coletadas 50.8% foram grau 5 na escala, já evidenciando de primeiro momento a aprovação da amostra de pesquisa. Considerando que na escala do grau 3 em diante representem os maiores níveis de necessidade de implementação temos um total de 82.7% das respostas. Esse resultado mostrado no gráfico da Figura 1.

Figura 1 – Pergunta 2



Fonte: Autoria própria.

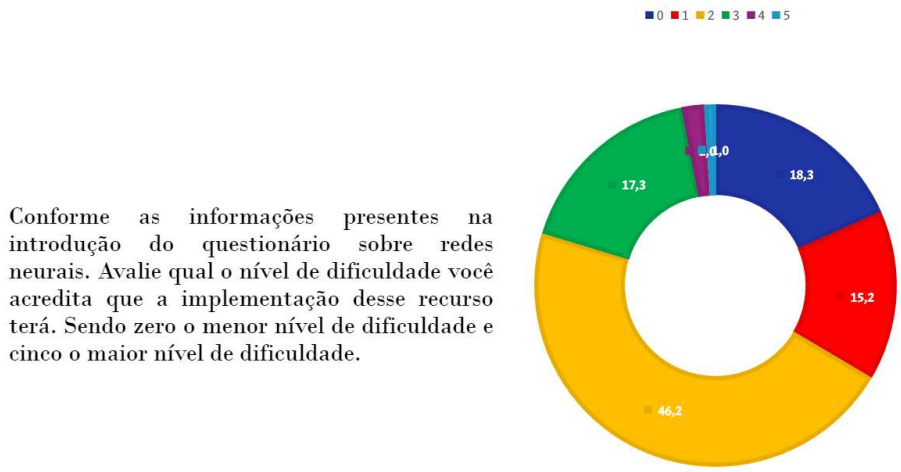
Na questão 6 foram expostas às informações sobre as ferramentas necessárias para implementação do projeto, colocando em foco que dessas ferramentas muitas já são conhecidas na área de segurança e que já são utilizadas por algumas empresas. Logo tendo em vista a informação exposta anteriormente o que resta para que o processo de implementação seja possível é o conhecimento necessário para tal. Assim a amostra foi indagada sobre o custo para implementação do modelo levando em conta que os profissionais envolvidos só precisaram obter o conhecimento para implementação. Obtendo como resultado um total de 80.7% de respostas "sim", definindo que os custos serão mínimos e apresentando um retorno positivo para mais uma questão chave da pesquisa. Os resultados podem ser vistos de forma gráfica na Figura 2.

Figura 2 – Pergunta 6

Fonte: Autoria própria.

Por fim na questão de número 10 foi indagado sobre o nível de dificuldade para implementação do modelo medido em uma escala de 0 a 5 onde quanto maior o número maior a dificuldade de implementação. Dentre as respostas obtidas vale ressaltar que a escala 0 que representa o menor grau de dificuldade para implementação do modelo obteve um total de 18.3% das respostas, a escala 1 obteve 15.2% das respostas e a escala 2 teve o maior número de respostas com um total de 46.2%. Sendo esses os menores números da escala que representa a dificuldade de implementação do trabalho, e sabendo que juntos representam um total de 79.7% de todas as respostas, podemos interpretar que a amostra acredita que o grau de dificuldade para implementação do modelo está entre médio e fácil, como mostra o gráfico presente na Figura 3.

Figura 3 – Pergunta 10

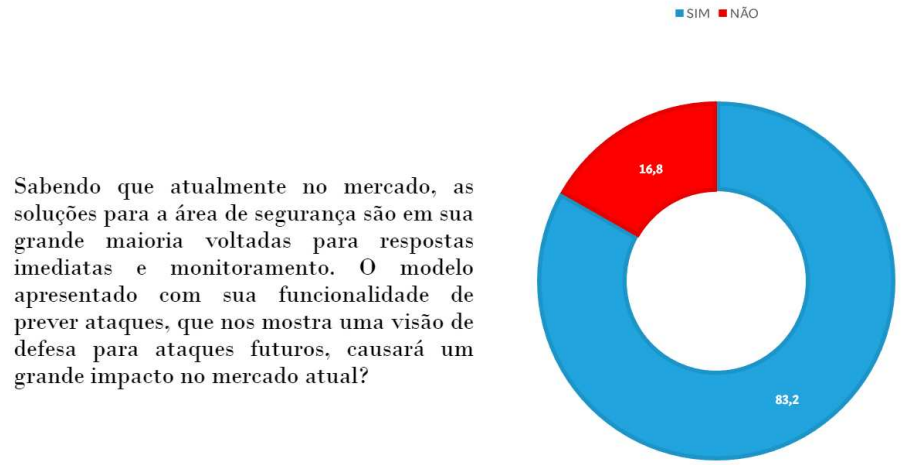


Fonte: Autoria própria.

4.2 Impacto causado pelo modelo

Nesta seção serão apresentados os dados coletados referentes aos possíveis impactos causados pela implementação do modelo.Para isso nas questões de números 1 e 3 foram coletados dados sobre o impacto causado no mercado atual após uma breve dissertação sobre o modelo. Na questão 1 os participantes responderam se haverá ou não impacto no mercado, onde foi obtido um total de 83.2% de respostas ”sim” representando 164 respostas, como demonstrado na Figura 4.

Figura 4 – Pergunta 1

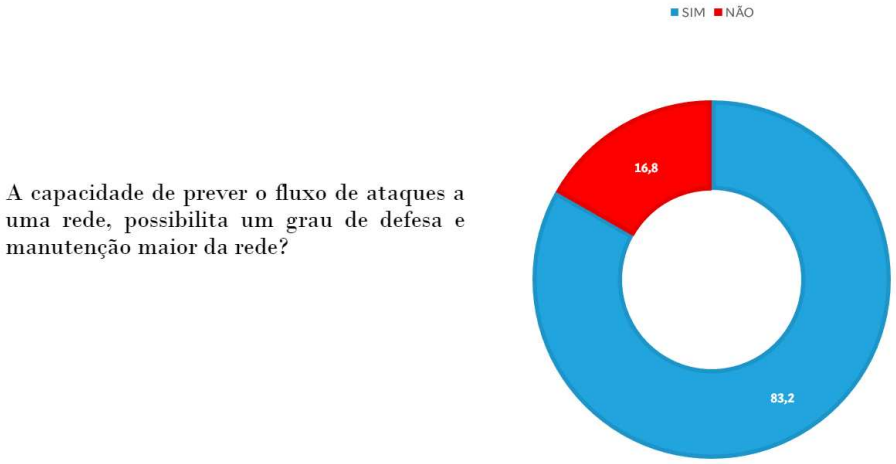


Fonte: Autoria própria.

Na questão de número 3 a pergunta foi referente a se o modelo após implementado poderá ofertar um grau maior de segurança de rede já que ele apresenta dados referentes a

futuros ataques. O resultado foi de 83.2% respostas "sim", ou seja um total 164 respostas, dados presentes no gráfico presente na Figura 5.

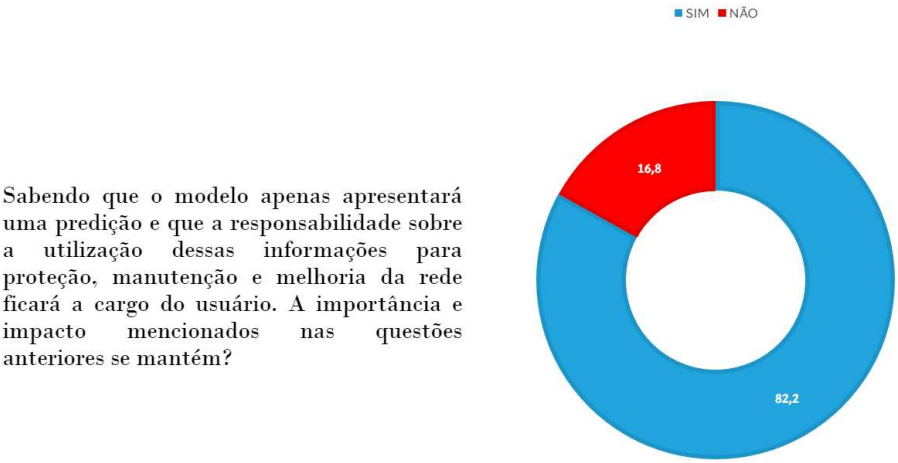
Figura 5 – Pergunta 3



Fonte: Autoria própria.

Nas questões de números 4 e 5 foram coletados dados sobre a valorização dos conhecimentos de cibersegurança e de redes neurais no mercado, assim como se o impacto causado se manteria após algumas considerações. Na questão 4 foi colocado que o modelo apenas apresentara dados sobre possíveis ataques e que a prevenção e tratativas deverão ser realizadas pelos profissionais da área de cibersegurança, com base nessa declaração foi questionado aos participantes impacto considerado nas questões anteriores se mantém, obtendo como retorno um total de 82.2% de respostas "sim", como presente no gráfico da Figura 6.

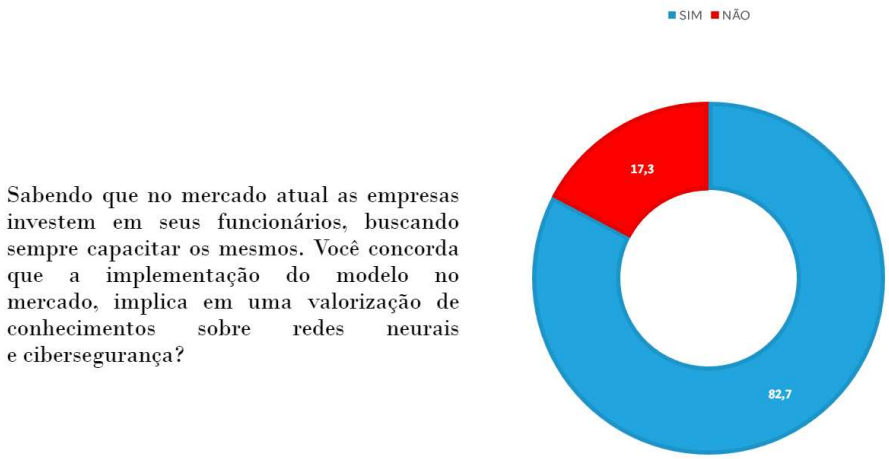
Figura 6 – Pergunta 4



Fonte: Autoria própria.

Por fim na questão 5 foi questionado se os conhecimentos, qualificação e treinamentos de profissionais nas áreas de cibersegurança e redes neurais aumentariam no mercado atual com a implementação desse modelo. O resultado foi de 82.7% de respostas "sim"representando 163 respostas, como mostra o gráfico da Figura 7.

Figura 7 – Pergunta 5



Fonte: Autoria própria.

4.3 Qualificação da Amostra

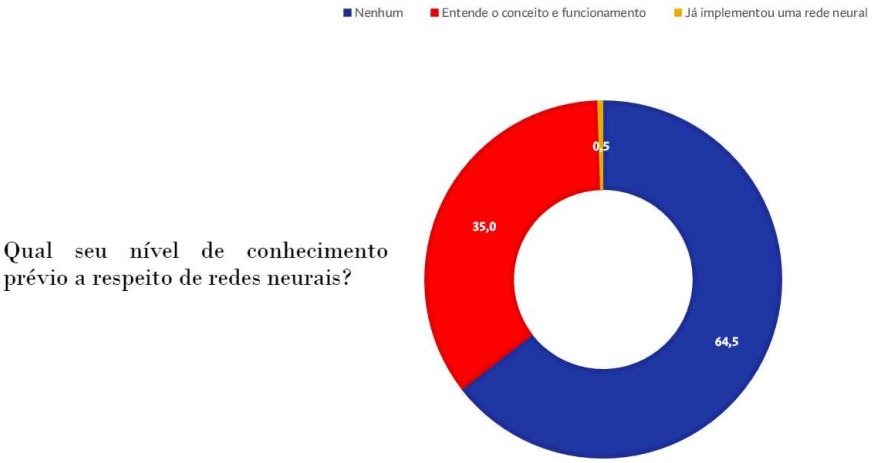
Nesta seção são apresentados os dados referentes as qualificações dos indivíduos da amostra, como conhecimentos prévios nas áreas relacionadas ao objeto de estudo e função atual na sociedade. Na questão 7 os participantes foram questionados sobre os seus conhecimentos prévios na área de cibersegurança. Para medir as respostas foi utilizado uma escala de 0 a 5 onde quanto menor o número menor o conhecimento sobre o tópico. Assim, o resultado da questão 7 foi de 45.7% para a escala 2 e se considerar que da escala 2 para baixo até a escala 0 o resultado é de 63.5% sendo 125 respostas. Com isso é possível considerar que o nível de conhecimento prévio para cibersegurança dos participantes se mantém entre médio e baixo. Como mostrado no gráfico da Figura 8.

Figura 8 – Pergunta 7

Fonte: Autoria própria.

Na questão 9 foi abordado os conhecimentos prévios para o tópico de redes neurais onde as respostas referentes ao nível de conhecimento são medidas em 3 níveis sendo eles: nenhum, entende o conceito e funcionamento, já implemento uma rede neural. Assim foi obtido como resultado um total de 64.5% para a resposta "nenhum", mostrando que 127 integrantes dos participantes não possuíam conhecimentos prévios sobre redes neurais. Como apresentado no gráfico presente na Figura 9.

Figura 9 – Pergunta 9



Fonte: Autoria própria.

Para categorizar os integrantes da amostra foi utilizada a questão número 8 onde foi perguntado qual a função atual dos participantes da amostra na sociedade. O resultado obtido foi de 47.2% dos entrevistados sendo estudantes, 19.8% sendo profissionais na área da tecnologia, 15.7% sendo profissionais em outras áreas e por fim um total de 17.3% sendo categorizados como ”outros” não se enquadrando assim a nenhuma das opções anteriores, veja a 10.

Figura 10 – Pergunta 8

Fonte: Autoria própria.

5 CONCLUSAO

Com base nos resultados do *survey*, pode-se concluir que houve uma resposta positiva por parte dos participantes. Os dados coletados revelaram uma tendência favorável em relação ao objeto de pesquisa. Os participantes expressaram opiniões e percepções positivas, o que indica um alto nível de aceitação para implementação do modelo de predição de ataques de rede utilizando *netflow*. Esses resultados podem refletir a viabilidade de implementação do modelo levando em consideração as dificuldades e custos, assim como a necessidade de implementação para com o mercado atual.

As respostas obtidas no *survey* sugerem que o objetivo da pesquisa para comprovar viabilidade e impacto estão alinhadas com as expectativas dos participantes. Essa avaliação favorável pode servir como um estímulo para implementar o modelo e continuar a realização de pesquisas em cima dele. É importante ressaltar que, apesar de nem todos os participantes possuírem conhecimentos prévios às áreas de correlatas ao modelo após explicações e situações propostas, os resultados obtidos foram positivos, denotando que até os participantes que não atuam na área de tecnologia reconheceram o impacto e viabilidade do modelo.

Em resumo, os resultados positivos obtidos pelo *survey* indicam uma boa aceitação ou aprovação com relação à viabilidade e impacto do modelo de predição de ataques de rede utilizando *netflow*. Isso fornece uma base sólida para continuar investindo no modelo que recebeu uma resposta positiva dos participantes.

REFERÊNCIAS

AGGARWAL, A. **Site sobre o elastic search**. 2022. Urlhttps://www.elastic.co/pt/what-is/elk-stack.

AGGARWAL, A. **Site sobre o elastic search**. 2022. Urlhttps://www.elastic.co/guide/en/beats/libbeat/current/beats-reference.htmlbeats-reference.

BAPTISTA, D. C. d. C. M. N. **Metodologias Pesquisa em Ciências - Análise Quantitativa e Qualitativa, 2ª edição**. Dissertação (Mestrado), 2016.

BOLFARINE, H.; BUSSAB, W. de O. **Elementos de amostragem**. [S.l.]: Editora Blucher, 2005.

CALIS, J. O. G. Modelo de predição de ataques de rede utilizando netflow. Universidade Estadual Paulista (UNESP), 2021.

CARVALHO, H. V. de; CARVALHO, E. C.; ARRUDA, H.; IMPERATRIZ-FONSECA, V.; SOUZA, P. de; PESSIN, G. Detecção de anomalias em comportamento de abelhas utilizando redes neurais recorrentes. In: SBC. **Anais do IX Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais**. [S.l.], 2018.

Cícero. **Questionario**. 2023. <https://docs.google.com/forms/d/e/1FAIpQLSc-oOMgC1R989qOSY1oojnxTlnsmhg8uySOsQeCmaETkiBsZw/viewform>, Último acesso: 07/07/2023.

FLáVIO, I. J. **PDF sobre SIMULAÇÃO POR EVENTOS DISCRETOS**. 2022. Urlhttps://cursos.unipampa.edu.br/cursos/engenhariadeproducao/files/2016/08/apostila-sim-simulacao-por-eventos-discretos.pdf.

FREITAS, H.; OLIVEIRA, M.; SACCOL, A. Z.; MOSCAROLA, J. O método de pesquisa survey. **Revista de Administra&ccdeil; ão da Universidade de São Paulo**, v. 35, n. 3, 2000.

GILBERT, M. da S.; CAMPOS, M. L. R. de; CAMPISTA, M. E. M. Autoencoders assimétricos para a compressao de dados iot. In: SBC. **Anais do XLI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**. [S.l.], 2023. p. 421–434.

GUPTA, Y. **Kibana essentials**. [S.l.]: Packt Publishing Ltd, 2015.

GUPTA, Y.; GUPTA, R. K. **Mastering Elastic Stack**. [S.l.]: Packt Publishing Ltd, 2017.

JURASZEK, G. D. *et al.* Reconhecimento de produtos por imagem utilizando palavras visuais e redes neurais convolucionais. Universidade do Estado de Santa Catarina, 2014.

KELLEHER, J. D. **Deep learning**. [S.l.]: MIT press, 2019.

KNUDSEN, L. R.; ROBSHAW, M. J. Brute force attacks. In: **The Block Cipher Companion**. [S.l.]: Springer, 2011. p. 95–108.

KUC, R.; ROGOZINSKI, M. **Elasticsearch server**. [S.l.]: Packt Publishing Ltd, 2013.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LYON, G. F. **Nmap network scanning: The official Nmap project guide to network discovery and security scanning**. [S.l.]: Insecure. Com LLC (US), 2008.

OLIVEIRA, R.; FERRO, M.; FERNANDO, B.; CARMELO, B. Avaliando técnicas de aprendizado profundo para detecção de esquistossomose mansoni em imagens de exames parasitológicos. **Proceedings of theCongresso Brasileiro de Inteligência Computacional**, 2018.

RAUBER, T. W. Redes neurais artificiais. **Universidade Federal do Espírito Santo**, v. 29, 2005.

SCHEUREN, F. What is a survey? In: AMERICAN STATISTICAL ASSOCIATION ALEXANDRIA. [S.l.], 2004.

SYSTEMS, I. A. r. r. C. **NetFlow Configuration Guide, Cisco IOS Release 15MT**. [S.l.]: Cisco Systems, Inc. (US), 2016.

VISSER, P. S.; KROSNICK, J. A.; LAVRAKAS, P. J. Survey research. Cambridge University Press, 2000.

WAINER, J. *et al.* Métodos de pesquisa quantitativa e qualitativa para a ciência da computação. **Atualização em informática**, Sociedade Brasileira de Computação/Editora PUC Rio Rio de Janeiro, v. 1, n. 221-262, p. 32–33, 2007.

ZUBA, M. V.; GOMES, R. M.; SANTOS, B. A. Análise de redes neurais adversariais generativas para a geração de imagens sint eticas. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 8, n. 1, 2021.