Revista Brasileira de Computação Aplicada, Julho, 2024

DOI: 10.5335/rbca.v16i2.15312
Vol. 16, Nº 2, pp. 16–30
Homepage: seer.upf.br/index.php/rbca/index

ARTIGO ORIGINAL

Identificando perfis de clientes de um ISP propensos ao *Churn* ao empregar dados do protocolo TR-069 nos algoritmos de ML

Identifying ISP's customers profiles prone to Churn by employing TR-069 protocol data into ML algorithms

Bernardo Gatto [1], Patricia Lengert [1], Augusto Abel [2], Yuri Bandeira [2], Raul Ceretta Nunes [3], Ricardo Tombesi Macedo [1]

[1]Universidade Federal de Santa Maria (UFSM), Campus Frederico Westphalen – Frederico Westphalen, RS – Brasil, [2]Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar), Campus Frederico Westphalen – Frederico Westphalen, RS – Brasil, [3]Universidade Federal de Santa Maria (UFSM), Campus Sede – Santa Maria, RS – Brasil

*{augusto.2022002506, yuri.2021002726}@aluno.iffar.edu.br;
†{bernardo.gatto, patricia.lengert}@acad.ufsm.br, {ceretta,rmacedo}@inf.ufsm.br

Recebido: 16/10/2023. **Revisado:** 07/07/2024. **Aceito:** 31/07/2024.

Resumo

Os provedores de serviço de Internet (*Internet Service Provider* – ISPs) oferecem uma infraestrutura de comunicação essencial para a realização de tarefas cotidianas da sociedade através da Internet e para o advento de novas tecnologias em rede. Todavia, um desafio dos ISPs consiste em reduzir a taxa de *churn*, a qual engloba a taxa de cancelamentos dos planos dos clientes. Apesar de existirem esforços na literatura, os ISP permanecem com uma carência de ferramentas para identificar o *churn*. Esse trabalho propõe o *ChurnSense*, um processo para identificar perfis de clientes de um ISP com base em técnicas de aprendizado de máquina, auxiliando no entendimento do problema do *churn*. O processo compreende três passos: Coleta, Pré-processamento e Análise. Por meio dele, a Coleta reúne os dados do TR-069, os quais são tratados pelo Pré-processamento e os perfis dos clientes são identificados pela Análise, fornecendo informações úteis na tomada de decisões sobre o *churn*. Um estudo de caso foi conduzido usando dados reais de um ISP regional. Os resultados obtidos mostram 20.61% dos dispositivos do clientes com qualidade de conexão aquém do desejado, estando em risco de *churn*.

Palavras-Chave: Perfis de Clientes; Protocolo TR-069; Provedor de Serviços de Internet.

Abstract

Internet Service Providers (ISPs) offer an essential communication infrastructure to support people everyday's tasks over the Internet and also the advent of new networking technologies. However, a challenge to ISPs is to reduce churn rate, also known as low customer retention rate. Despite efforts in the literature, ISPs remain short of tools to identify customers' churns. This paper proposes ChurnSense, a process to identify ISP customers profiles by employing machine learning techniques, contributing to the understanding of the churn problem. The processes comprises three steps: Collect, Pre-processing, and Analysis. Through it, Collect gathers data from TR-069 protocol, Pre-processing treats these data and Analysis finds clusters that define customers profiles, providing useful information to decision making about churn. A case study was conducted by employing real data from a regional ISP. The obtained results show 20.61% of customers devices with connection quality below the expected, being at risk of churn.

Keywords: Customers Profiles; Internet Service Provider; TR-069 Protocol.

1 Introdução

A Internet desempenha um papel fundamental na sociedade atual, proporcionando acesso aos serviços e aplicações de diferentes áreas, tais como saúde, entretenimento, educação e segurança (Utamima et al., 2023). Em decorrência da alta demanda de acesso destes serviços, os provedores de serviço de Internet (ISPs - *Internet Service Providers*) surgem como instituições que visam suprir estas necessidades ao proporcionar uma experiência diferenciada aos seus clientes (Servio et al., 2023). Além disso, os ISPs são estratégicos para o desenvolvimento tecnológico, fornecendo uma infraestrutura para dar suporte ao surgimento de novas tecnologias em rede, tais como, a Internet das Coisas (IoT – *Internet of Things*) (Mostacero-Agama e Shiguihara, 2022) e a sexta geração de telefonia móvel (6G) (Sambhwani et al., 2022).

Todavia, o *churn* representa um desafio para assegurar a viabilidade financeira dos ISPs (Pebrianti et al., 2022). O modelo de negócio de um ISP se baseia na subscrição de consumidores aos seus planos de Internet (Ikhsan et al., 2022). Quando um consumidor decide se desinscrever do plano contratado, ocorre o fenômeno conhecido como *churn* (Dai et al., 2021). Em linhas gerais, o *churn* consiste no cancelamento do plano contratado de um ISP, seguido de uma eventual subscrição em um provedor concorrente (Prasetyo et al., 2022). O *churn* representa um desafio real para os ISPs devido ao aumento de empresas que prestam serviços de Internet e à competitividade natural por clientes com base na oferta de uma melhor qualidade de serviço com um menor preço (Peddarapu et al., 2022). Em decorrência do *churn*, surge a necessidade da elaboração de estratégias para compreender este fenômeno.

Na literatura, os principais trabalhos podem ser classificados em três linhas de pesquisa. A primeira busca empregar o protocolo TR-069 (Forum, 2020) para obter dados mais precisos da rede doméstica, tais como o indicador de intensidade do sinal recebido (*Received Signal Strength Indication - RSSI*) (Lygerou et al., 2022). A segunda investiga o uso das séries temporais para melhor compreender os perfis dos clientes do ISP (Santos et al., 2019). A terceira aborda técnicas de aprendizado de máquina (*Machine Learning* – ML) para identificar grupos de perfis de consumidores em outros contextos, tais como nas redes inteligentes de energia elétrica (Wang et al., 2022). No entanto, mesmo existindo estes esforços, ainda existe uma lacuna no estado da arte e os ISP permanecem carentes de metodologias capazes de especificar como os dados oriundos das redes domésticas dos clientes podem ser analisados de forma a gerar subsídios para identificar eventuais cancelamentos dos clientes.

Este trabalho apresenta um processo para identificação dos perfis de clientes de um ISP ao aplicar técnicas de ML, contribuindo para o diagnóstico do problema do *churn*. O processo segue três passos, a Coleta, o Pré-processamento e a Análise. A Coleta reúne dados obtidos dos roteadores instalados nos clientes por meio do protocolo TR-069. O Pré-processamento trata e refina os dados coletados. A Análise aplica algoritmos de clusterização para identificar perfis de clientes com base na análise da intensidade do sinal dos dispositivos conectados aos roteadores. Na análise, o RSSI consiste em uma das métricas de interesse, pois pode ser coletado por meio do protocolo TR-069, recebendo atenção especial no processo proposto para representar a qualidade do sinal entre o dispositivo sem fio da casa do cliente e o seu ponto de acesso à rede. Com base no resultado da análise (perfis identificados), o ISP pode elaborar planos de ação para aprimorar a qualidade do serviço prestado aos seus clientes e potencializar a redução de *churns*.

Um estudo de caso foi conduzido utilizando dados reais provenientes de um ISP regional. Estes dados foram coletados de mais de três mil roteadores residenciais espalhados pela rede do ISP em um período de quinze dias. Os critérios observados na análise consistiram na quantidade de dispositivos conectados na rede doméstica e no RSSI de cada dispositivo pertencente à rede.

Os resultados obtidos nessa análise revelam um perfil de clientes específico, abrangendo aproximadamente 20.61% dos dispositivos pertencentes aos clientes, com indícios significativos de um potencial cancelamento dos serviços (*churn*). Esses dados são extremamente relevantes para a compreensão do comportamento dos clientes e podem fornecer *insights* valiosos para estratégias de retenção e melhorias no atendimento.

O trabalho está organizado como segue. A Seção 2 explica os conceitos fundamentais do artigo. A Seção 3 apresenta os trabalhos relacionados. A Seção 4 descreve o processo proposto para identificar os perfis dos clientes de um ISP usando ML e aprendizado de máquina. A Seção 5 detalha os resultados obtidos. A Seção 6 discute como os resultados obtidos podem ser empregados pelos ISP com o objetivo de mitigar o *churn*. Enquanto que a Seção 7 conclui o trabalho.

2 Fundamentação

Esta seção apresenta os fundamentos teóricos necessários para o entendimento do artigo. A Seção 2.1 descreve o protocolo TR-069, a Seção 2.2 detalha as abordagens de ML e a Seção 2.3 apresenta dois algoritmos de ML com potencial de emprego em ISPs.

2.1 O Protocolo TR-069

O protocolo TR-069 (Forum, 2020), também conhecido como *CPE WAN Management Protocol* (CWMP), consiste em um protocolo de gerenciamento remoto desenvolvido para dispositivos de rede residenciais e corporativos, conhecidos como *Customer Premises Equipment* (CPEs). O TR-069 permite que *Internet Service Providers* (ISPs) gerenciem e configurem esses dispositivos remotamente, facilitando a instalação, o monitoramento e a solução de problemas.

Uma das principais vantagens do TR-069 consiste na capacidade de gerenciar eficientemente uma grande quantidade de dispositivos CPEs distribuídos geograficamente. Isso permite que os ISPs automatizem tarefas de rotina, reduzam o tempo e os custos associados ao suporte técnico e melhorem a qualidade do serviço oferecido aos assinantes. Devido a isso, os ISPs passaram a considerar uma prioridade a necessidade de gerenciar e configurar eficientemente a crescente quantidade de equipamentos.

O protocolo TR-069 define uma série de comandos e

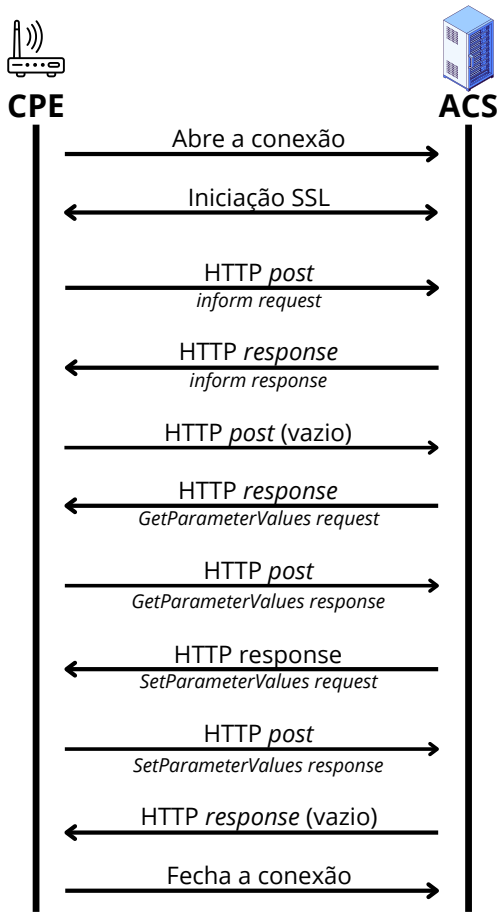

Figura 1: Fluxo de comunicação HTTP TR-069

parâmetros que permitem ao ACS (*Auto Configuration Server*) controlar e configurar remotamente o CPE. Isso inclui a instalação e atualização de *firmware*, configuração de parâmetros de rede, monitoramento de desempenho e obtenção de informações sobre o estado do dispositivo. O protocolo também fornece recursos para a solução de problemas, como a reinicialização remota e a coleta de registros de eventos.

O funcionamento do protocolo TR-069 (Huang, 2019) tem como base uma arquitetura composta por dois componentes principais: o Servidor ACS e o Agente de CPE. O ACS, fornecido pelo ISP, atua como o ponto central de gerenciamento, enquanto o Agente de CPE compreende os dispositivos a serem gerenciados. Essa arquitetura adota um modelo de comunicação cliente-servidor, onde o CPE age como o cliente e o ACS como servidor. A comunicação ocorre através de conexões via HTTP (*Hypertext Transfer Protocol*), tornando o protocolo TR-069 uma solução eficiente para a configuração inicial e o gerenciamento contínuo de dispositivos. Com base na Fig. 1, durante a comunicação entre o ACS e a CPE, o ACS estabelece uma conexão, envia solicitações HTTP para obter informações e configurar parâmetros, e recebe respostas da CPE confirmando ações. Esse processo permite o gerenciamento remoto da CPE. Após a conclusão da troca de informações, a conexão encerra.

Os servidores ACS desempenham um papel importante no cenário de redes de comunicação modernas, especialmente em um contexto de rápida expansão de dispositivos conectados à Internet (Basicevic, 2023). Nesse contexto os Servidores ACS oferecem uma plataforma versátil capaz de lidar com essa ampla gama de dispositivos e protocolos de comunicação em um único ponto central. Eles fazem um trabalho fundamental de gerenciamento de dispositivos na rede, permitindo que os administradores realizem uma série de tarefas cruciais. Algumas destas tarefas consistem na definição dos servidores DNS (*Domain Name System*), a atualização de *firmware* e o monitoramento de *status* da conexão. Diante dos benefícios proporcionados pelo protocolo TR-069, existe um movimento entre os fabricantes para disponibilizar nativamente este protocolo em diversos modelos de dispositivos para ISPs (Vargas Maquilon e Maruri Uriña, 2021).

A Fig. 2 ilustra os principais elementos de parâmetros do TR-069 que estão organizados em forma de árvore. Os nós folhas representam os valores do parâmetros finais e definem uma configuração específica para controlar o comportamento e as funcionalidades dos dispositivos gerenciados remotamente. Nessa estrutura, o nó raiz compreende no parâmetro *InternetGatewayDevice*, que representa o *gateway* de Internet ou roteador. Os nós filhos se subdividem em *DeviceInfo*, *LANDevice* e *WANDevice*. Dentro do *DeviceInfo* são disponibilizados parâmetros que descrevem os detalhes e as características essenciais do dispositivo como modelo, fabricante, versão de *firmware*. O *LANDevice*, consiste na subárvore mais rica em atributos, abordando a rede local do dispositivo e as interfaces LAN, configurações de IP (*Internet Protocol*), DHCP (*Dynamic Host Configuration Protocol*). Enquanto que o *WANDevice* agrega atributos para descrever a rede de longa distância do dispositivo e detalhes sobre configurações de conexão, endereços IP e informações de roteamento.

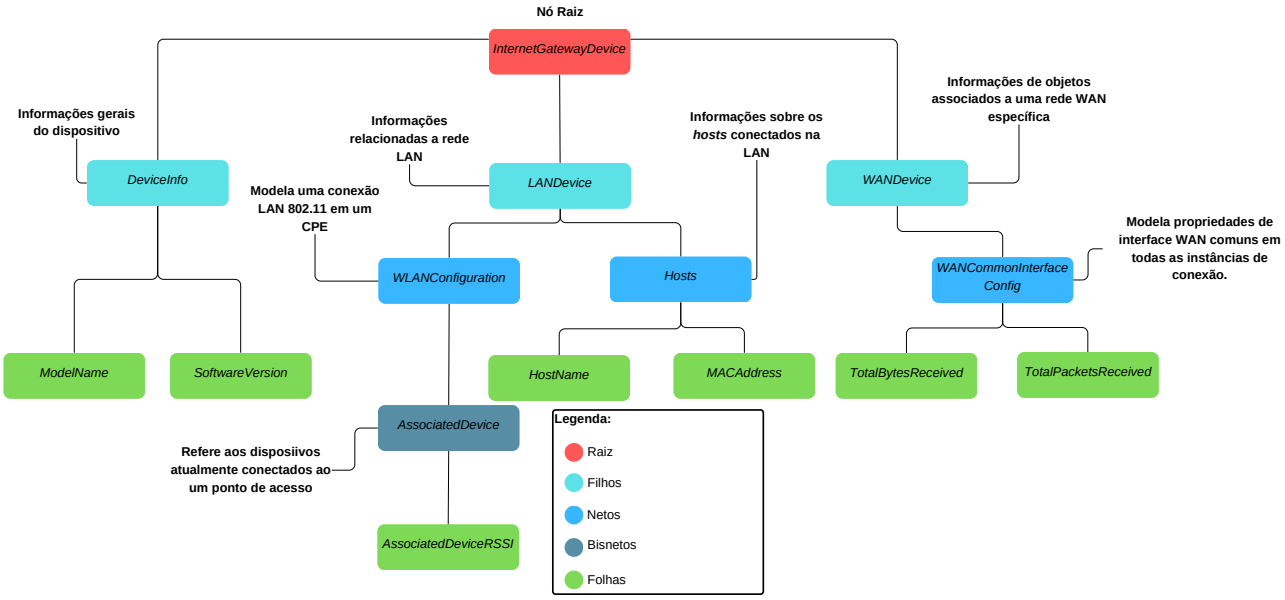

Figura 2: Parâmetros do Protocolo TR-069

Dentre estes parâmetros, o RSSI desempenha um papel estratégico no monitoramento. O RSSI faz parte dos parâmetros essenciais em dispositivos sem fio, como roteadores Wi-Fi, *smartphones* e aparelhos *bluetooth*, medindo a intensidade do sinal recebido por radiofrequência, expresso em dBm (decibéis na escala *miliwatts*). Essa informação serve para determinar a qualidade da conexão sem fio e a força do sinal captado pelo dispositivo receptor. Ao integrar o RSSI e outros parâmetros relevantes, os servidores ACS se tornam capazes de fornecer um panorama completo do estado e desempenho dos dispositivos conectados para os ISPs, tornando possível o gerenciamento e o monitoramento eficiente de toda a rede.

Com o protocolo TR-069, várias operações podem ser realizadas pelo ISP, reduzindo seus custos de atendimentos presenciais. Alguns exemplos destas operações, consistem na configuração e no monitoramento de parâmetros do dispositivo, atualização de *firmware*, diagnóstico remoto, provisionamento de serviços, sendo a configuração do CPE do cliente. Logo, essas operações permitem que os ISPs e administradores de rede gerenciem eficientemente uma ampla gama de dispositivos de forma remota, garantindo maior controle, resolução de problemas mais rápida e redução de custos operacionais.

2.2 Abordagens de ML

O ML engloba uma área da ciência da computação que permite ao computador aprender automaticamente, sem ser programado explicitamente. Ele evoluiu a partir do reconhecimento de padrões, tendo como foco fazer previsões ou agrupamentos com base em dados e estatísticas computacionais (N e Gupta, 2020).

Os algoritmos de ML permitem analisar dados e criar modelos de predição precisos. Eles contribuem para melhorar o desempenho das redes, prever falhas de equipamentos e otimizar o uso de recursos (Campanile et al., 2021). As principais abordagens de ML incluem aprendizagem supervisionada, não supervisionada, semi supervisionada e por reforço (N e Gupta, 2020), onde cada uma apresenta suas próprias vantagens e aplicações. Em resumo, os algoritmos de ML representam uma ferramenta valiosa para ISPs, permitindo-lhes melhorar o desempenho e a experiência do cliente.

A abordagem supervisionada depende de um conjunto de dados de treinamento com rótulos. Em linhas gerais, um rótulo consiste na característica principal a ser analisada que está contida em uma amostra de dados (Otchere et al., 2021). Os algoritmos desta abordagem aprendem a partir dos rótulos para fazer previsões ou tomar decisões sobre novos dados. A abordagem tem como objetivo construir um modelo para mapear as características dos dados de entrada para os rótulos correspondentes.

A abordagem de ML não supervisionada não necessita de rótulos nos dados de treinamento. Eles exploram a estrutura intrínseca dos dados para identificar padrões, agrupamentos ou relações ocultas. Esses algoritmos se mostram úteis para descobrir informações sobre os dados sem a necessidade de conhecimento prévio dos rótulos (Usama et al., 2019).

O aprendizado semi supervisionado combina elementos do aprendizado supervisionado e do aprendizado não su-

pervisionado. Ao contrário do aprendizado supervisionado, onde cada exemplo de treinamento está obrigatoriamente rotulado com uma classe, e do aprendizado não supervisionado, em que não há rótulos disponíveis, o aprendizado semi supervisionado trabalha com um conjunto de dados parcialmente rotulado (N e Gupta, 2020).

O aprendizado por reforço consiste em um paradigma de ML inspirado pela forma como os seres humanos aprendem por meio de tentativa e erro, com base em recompensas e punições. Nesse tipo de aprendizado, um agente de aprendizado interage com um ambiente dinâmico, tomando ações e recebendo *feedbacks* na forma de recompensas ou penalidades, que indicam o desempenho do agente em relação a um objetivo específico (Zhao et al., 2020).

Conhecida as abordagens de ML, cabe destacar que para a identificação dos perfis de clientes de um ISP a partir da análise de dados coletados em roteadores que executam o protocolo TR-069, não foram encontrados dados rotulados que permitam aprendizado satisfatório. Por outro lado, observa-se que técnicas não supervisionadas são adequadas para agrupamento de dados e redução de dimensionalidade, atendendo as necessidades deste trabalho. Tais propriedades auxiliam na análise comportamental de clientes de ISP e na identificação do perfil dos clientes propensos a Churn.

2.3 Algoritmos de ML e ISPs

Os algoritmos K-Means e a Rede Neural SOM (*Self-Organizing Map*) têm se mostrado instrumentos valiosos e de grande interesse para as redes e ISPs (Yang e Hussain, 2023; Liao et al., 2022). O K-Means proporciona um agrupamento eficiente de dispositivos, permitindo a otimização dos recursos da rede (Yang e Hussain, 2023). A Rede Neural SOM mostra utilidade para analisar séries temporais e detectar padrões, contribuindo para a eficiência e desempenho das redes (Liao et al., 2022).

Classificados na abordagem não supervisionada, os algoritmos K-Means e a Rede Neural SOM possuem propriedades interessantes. O algoritmo K-Means permite agrupar os dados em k grupos distintos, onde k representa um número pré-definido para os grupos que serão criados (Bajpai e He, 2020). Ele também possui eficiência computacional e escalabilidade. Sua implementação apresenta simplicidade e possui um tempo de convergência rápido, tornando-o adequado para aplicações em tempo real ou conjuntos de dados grandes, como no caso dos ISPs. O objetivo do K-Means consiste em centralizar o máximo possível os centroides com base na distância entre os pontos e o centro do agrupamento, como ilustrado na Fig. 3. O processo de utilização do algoritmo é simples, tal como segue: *i*) definição do número de centroides ou *clusters*; *ii*) treinamento e atualização dos centroides em X etapas; *iii*) resultado final do agrupamento. A Fig. 3 mostra um exemplo do passo a passo da utilização do K-Means. Suponha que os pontos verdes identificam dispositivos como celulares ou computadores com RSSI bom, e os pontos em laranja representam dispositivos com o RSSI suficiente, como fica nítido a separação dos dados linearmente, o número de *clusters* foi definido como dois. Na primeira iteração (Treino 1), os centros renovam sua posição através do cálculo da distância euclidiana, onde o centroide

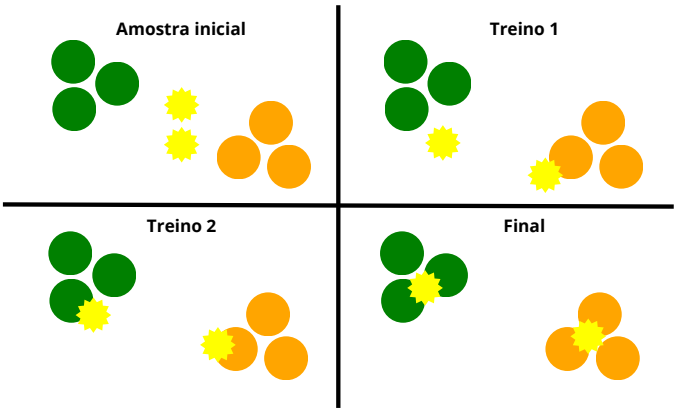

Figura 3: Treinamento do K-Means

encontra-se "puxado" para a média, esse cálculo então se realiza novamente, porém com os dados atualizados de cada iteração sendo repetidos até que os centroides se estabilizem, e não haja mais mudanças significativas nas atualizações, no final o resultado obtido neste exemplo consiste em dois *clusters*, verde (RSSI bom) e laranja (RSSI suficiente), com um centroide e pontos (dados) a ele atribuídos.

Por se tratar de um tipo de rede neural, a SOM herda características intrínsecas desta abordagem de aprendizagem de máquina. Dentre estas características, cabe mencionar o emprego de neurônios, dois estágios distintos de operação e o uso de camadas para resolução de problemas de clusterização (Mingoti e Lima, 2006). Em relação ao uso de neurônios, a rede neural SOM ajusta os pesos de entradas dos neurônios e emprega o modelo de Perceptron. Além disso, a rede neural SOM utiliza os estágios de treinamento e o estágio de recall para classificar as amostras. Assim como nas redes neurais tradicionais, a SOM emprega o conceito de camadas de entrada, camadas ocultas e uma camada de saída para representar os resultados.

Todavia, a principal especificidade da rede neural SOM consiste no emprego da camada *Kohonen*, cujo nome consiste em uma homenagem ao criador da SOM (Li e Zhu, 2018). A camada *Kohonen* usualmente é projetada como um arranjo bidimensional de neurônios que mapeiam entradas com n dimensões para apenas duas dimensões. Um importante ponto neste mapeamento consiste na preservação das características da topologia do espaço de entrada para os *clusters* identificados.

Os parâmetros da rede neural SOM podem ser melhor explicados de forma matemática. A rede neural consiste em um *array* $M = m \times m$ formada por neurônios de processamento. Se estes neurônios estão organizados em um plano em forma de *grid*, logo, podemos afirmar que esta rede é bidimensional, pois esta rede mapeia vetores de entrada com n dimensões para uma plano com apenas duas dimensões. Para uma dada rede, o vetor de entrada x possui uma dimensão fixa n. Os n componentes do vetor de entrada x (isto é, $x_1, x_2, ..., x_n$) estão conectados com cada neurônio no *array*. Um peso sináptico w_{ij} é definido como uma conexão para o $i-^{simo}$ componente do vetor de entrada para o $j-^{simo}$ neurônio. Sendo assim, um vetor n-dimensional w_j de pesos sinápticos está associado com

cada neurônio *j*.

As duas questões centrais para um algoritmo de aprendizado de uma rede neural SOM são: *i)* o peso do processo de adaptação e *ii)* a ideia de uma topologia da vizinhança dos neurônios. Estas redes operam em duas fases, sendo elas a busca por similaridades e a fase de adaptação dos pesos. Inicialmente, os pesos são configurados para valores aleatórios pequenos e um padrão é apresentado para o conjunto de nós de entrada da rede. Durante a fase de busca por similaridades, são computadas as distâncias Euclidianas entre as entradas e os pesos associados com os neurônios de saída. Depois, é escolhido como vencedor, um neurônio de saída *j* com a menor distância entre os *M* neurônios de saída. Na segunda fase, os pesos dos nós de entrada para o nó vencedor são modificados. Além disso, é identificada a topologia da vizinhança do nó vencedor e os pesos que convergem para estes neurônios são modificados.

Dentre os modelos estudados, o modelo K-Means e a Rede Neural SOM possuem características que os diferenciam em relação aos demais. Por operar de maneira eficiente com grandes quantidades de dados, o modelo K-Means vêm se mostrando como uma abordagem popular para a clusterização (Ikotun et al., 2023). Enquanto que a Rede Neural SOM oferece a possibilidade de evitar que o centro da clusterização compreenda uma solução ótima local (Zhu e Han, 2024). Portanto, como os dados coletados por meio do protocolo TR-069 podem compreender grandes volumes de dados, o K-Means apresenta um potencial para este cenário e a sua combinação com a neural SOM permite a busca por soluções ótimas globais para o centro dos clusters formados para representar os perfis dos clientes.

3 Trabalhos Relacionados

Os trabalhos do estado da arte podem ser organizado em quatro grupos: *i)* voltados ao emprego do protocolo TR-069; *ii)* voltados ao uso de técnicas tradicionais de séries temporais em dados de ISP, incluindo a identificação de perfis de usuários; *iii)* voltados a aplicar algoritmos de ML essencialmente em dados cadastrais de clientes de ISPs para predizer o *churn*; e *iv)* voltados a utilização de algoritmos de clusterização em séries temporais, mas em outros contextos. A seguir são apresentados trabalhos de cada um destes grupos.

O protocolo TR-069 está despertando o interesse de muitos pesquisadores ao facilitar o provisionamento e gestão dos dispositivos dos clientes do ISP (*Grupo 1*) ao possibilitar a coleta do número de dispositivos conectados por roteador e o RSSI de cada dispositivo na rede doméstica. Hils e Böhme (2020) apresentaram uma pesquisa que mostra a aderência do protocolo TR-069 nos ISP europeus, salientando boas práticas de segurança na utilização deste protocolo. Liao e Wang (2022) propuseram uma extensão da arquitetura do protocolo TR-069 para torná-lo descentralizado com base na tecnologia de *blockchain*. Lygerou et al. (2022) apresentaram uma estratégia de *honeypot* construída com base no protocolo TR-069 para dispositivos da Internet das Coisas. No entanto, nativamente, este protocolo não prevê o emprego de técnicas sofisticadas para melhor compreender o comportamento dos clientes, apenas viabiliza o monitoramento.

Diferentes autores propuseram analisar dados coletados de um ISP com suporte da abordagem das séries temporais (*Grupo 2*). Streit et al. (2019) aplicaram séries temporais com o objetivo de entender o perfil do comportamento dos clientes do ISP ao analisar o tráfego de *download* e *upload* dos roteadores. Ximenes et al. (2018) apresentaram uma abordagem para prever congestionamentos transitórios nas redes dos ISPs ao aplicar séries temporais em dados coletados a partir de comandos *traceroutes*. Santos et al. (2019) propuseram um método para detectar anomalias baseado em um modelo estatístico ao correlacionar medições de QoS e chamados de *call center* do ISP. Todavia, os dados empregados nestas analises se restringem aos dados de *download* e *upload* dos clientes, dando pouca atenção ao RSSI e ao número de dispositivos conectados por roteador, não sendo efetivo na mitigação de *churns*.

Outros trabalhos empregam algoritmos de aprendizado de máquina para identificar agrupamentos de dados em outros contextos (*Grupo 3*). Wang et al. (2022) propuseram uma abordagem de aprendizado federada para agrupar perfis de consumidores de eletricidade usando o algoritmo K-Means. Este algoritmo está recebendo atenção dos pesquisadores devido a sua boa capacidade de segmentar grupos de clientes (Nandapala e Jayasena, 2020). Tan et al. (2022) empregaram a rede neural SOM *(Self Organizing Map)*, um algoritmo de aprendizado não supervisionado que define um mapeamento de um espaço dimensional contínuo, com o objetivo de alocar múltiplas tarefas para SUV (*Unmanned Surface Vehicles*). No entanto, estas análises não consideraram clientes dos ISP.

Outros estudos (*Grupo 4*) investigaram como aplicar algoritmos de ML essencialmente em dados cadastrais dos clientes de ISPs como uma alternativa para prever o *churn*. Jain et al. (2020) apresentaram um modelo para predizer o *churn* de clientes usando dois algoritmos de ML, a regressão logística e a *Logit Boost* e para realizar experimentos usaram dados obtidos da Orange, uma companhia norte-americana de telecomunicações. Considerando os dados de um ISP da nação caribenha de Trindade e Tobago, Bachan e Gaber (2021) propuseram um método de previsão do *churn* de clientes que emprega os algoritmos de ML de árvore de decisão, regressão logística e máquina de vetores de suporte. Rahman et al. (2022) advogaram sobre a proposição de modelo usando três estágios: coleta de dados, identificação de valores nulos e processamento de dados, onde os algoritmos de redes neurais, regressão logística e o algoritmo de aprendizado de impacto foram comparados quanto a predição da ocorrência do *churn* de clientes. Entretanto, estes estudos assumem a existência de uma base de dados sobre os serviços contratados pelos consumidores e as suas respectivas informações cadastrais, desconsiderando os dados capazes de descrever como estava a qualidade dos serviços ofertados pelo plano de dados, tais como, dados que podem ser obtidos do protocolo TR-069, bem como o emprego destes dados para classificar diferentes perfis de clientes e identificar clientes propensos ao *churn*.

Este trabalho contribui para o estado da arte ao apresentar um processo de análise para identificar perfis do cliente que contempla o protocolo TR-069, empregando técnicas de séries temporais e algoritmos de ML. Assim como os trabalhos do *Grupo 4*, este trabalho aborda como aplicar

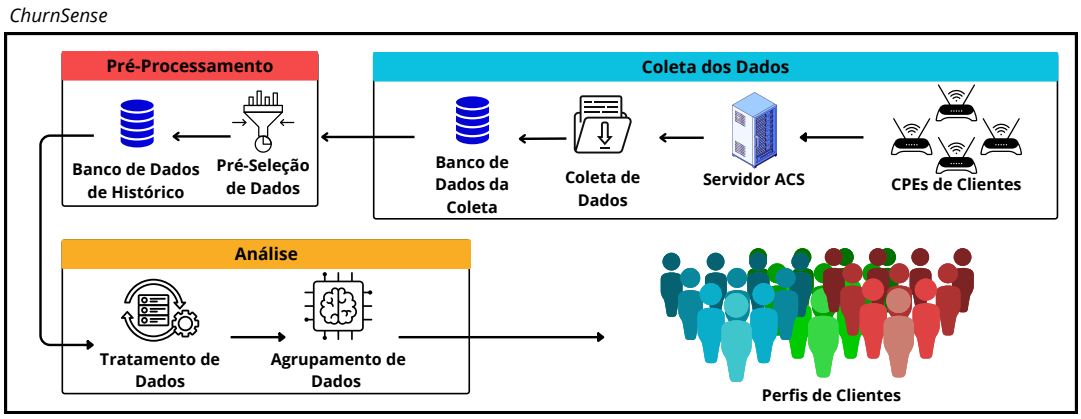

Figura 4: Diagrama do Processo *ChurnSense*

algortimos de ML para o problema do *churn*, mas considerando dados que retratam a qualidade do serviço prestado na residência dos consumidores. Além disto, da mesma forma que os estudos do *Grupo 1*, este processo considera o protocolo TR-069, mas utilizando os algoritmos K-Means e SOM oriundos do *Grupo 4* e mostrando seu impacto nas técnicas de séries temporais oriundas do *Grupo 2*.

4 *ChurnSense*: Identificando Perfis de Clientes de um ISP Propensos ao *Churn*

Esta seção apresenta o *ChurnSense*, um processo para identificar os perfis dos clientes do ISP com maior propensão ao *churn*, tendo como suporte dados coletados pelo protocolo TR-069. O processo compreende três passos (vide Fig. 4): a Coleta, o Pré-processamento e a Análise. A Coleta reúne os principais dados dos roteadores dos clientes, como por exemplo, o número de dispositivos conectados e o RSSI de cada dispositivo em um instante de tempo, e armazena no Banco de Dados da Coleta. O Pré-Processamento trata os dados provenientes da coleta, a fim de filtrar os dados mais relevantes para o uso na próxima etapa, armazenando os dados organizados no Banco de Dados de Histórico. Por fim, a etapa de Análise emprega os algoritmos de ML para agrupar os dados pré-processados e indicar os perfis encontrados. Em síntese, a partir de dados do protocolo TR-069 o processo proposto identifica o perfil com maior tendência ao *churn* e com base nos perfis identificados um ISP pode realizar ações específicas de retenção de clientes, como oferecer *upgrades* de serviço, melhorar a qualidade do sinal e fornecer suporte técnico especializado. As seções seguintes detalham cada passo do processo *ChurnSense*.

4.1 Coleta

A coleta dos dados alicerça todo o desenvolvimento do processo *ChurnSense*, pois ela proporciona os dados brutos utilizados para alimentar os algoritmos de ML (Banco de Dados da Coleta), coletados com base no protocolo de rede TR-069.

Usando o protocolo TR-069, o passo da Coleta obtém os dados dos roteadores de forma remota por meio de um servidor ACS. Para que as informações dos roteadores sejam armazenadas no servidor ACS, assumimos uma premissa de que todos os roteadores do ISP estão provisionados. Com base nesta premissa, este passo opera com o suporte de um *inform*, que consiste em uma ação vinda do roteador (cliente) para comunicar ao ACS (servidor), enviando as principais informações contidas no roteador para que o ISP tenha acesso. A partir disso, o servidor ACS se responsabiliza por mostrar os dados dos roteadores de clientes do ISP, tendo a capacidade de gerenciar remotamente a rede do provedor e fornecer dados estratégicos sobre a rede interna do cliente.

O Banco de Dados da Coleta no contexto do processo *ChurnSense* desempenha um papel crucial ao armazenar as informações coletadas na interface do servidor ACS. Dois requisitos expõem características essenciais para sua eficácia. O primeiro consiste na capacidade de armazenar a maior quantidade possível de informações brutas dos roteadores domésticos, incluindo os parâmetros relevantes. O segundo requisito determina que os dados dos roteadores sejam organizados de uma forma que possam ser usados em séries temporais. Essa coleta e armazenamento adequados garantem uma análise detalhada dos dados, proporcionando *insights* importantes para o entendimento do *churn* e o aprimoramento contínuo da rede.

4.2 Pré-Processamento

O Pré-Processamento compreende duas etapas cruciais do processo *ChurnSense*: a Pré-Seleção e o armazenamento no Banco de Dados de Histórico. A etapa da Pré-Seleção dos Dados faz parte da identificação das variáveis mais importantes para o modelo a ser utilizado. Nesta seleção as variáveis com maior representatividade para o processo *ChurnSense* consistem no RSSI e na quantidade de dispositivos conectados. Essa seleção criteriosa tem como orientação a premissa de preservar apenas os dados úteis, simplificando a análise subsequente e tornando-a mais eficiente. Os dados filtrados são então devidamente armazenados no Banco de Dados de Histórico, projetado para armazenar os dados por um período maior de tempo, sendo uma valiosa fonte para análises retrospectivas e projeções futuras, possibilitando a identificação de tendências e padrões ao

longo do tempo.

4.3 Análise

O passo de Análise recebe como entrada os dados vindos do passo de Tratamento de Dados e na sequência aplica estes dados em algoritmos de ML. Antes de aplicar os algoritmos de agrupamento, o processo *ChurnSense* necessita predeterminar quantos perfis foram gerados. Desta forma, o Método do Cotovelo (*Elbow Method*) é aplicado para auxiliar na determinação do número ideal de agrupamentos ou *clusters*. O processo *ChurnSense* emprega o Método do Cotovelo em vez de outros métodos, como o coeficiente de silhueta, o índice de Calinski-Harabasz ou a estatística *Gap*, devido à sua facilidade de uso e ampla aceitação e utilização na comunidade científica (Sarjonen e Höyhtyä, 2023).

Para determinar os perfis, o processo *ChurnSense* emprega dois algoritmos não supervisionados, ou seja, os dados não possuem um rótulo ao qual o algoritmo possa recorrer e conferir os resultados encontrados. O processo *ChurnSense* emprega o algoritmo K-Means e a Rede Neural SOM para descobrir perfis de clientes, sendo que cada um deles possui um objetivo específico. O uso do algoritmo K-Means identifica o perfil de clientes propensos ao *churn* proporcionando uma compreensão detalhada dentro do intervalo de um único dia. Já o emprego da Rede Neural SOM visa complementar a identificação ao permitir uma análise dos perfis ao longo de um intervalo de tempo maior, dedicando-se a análise da série temporal como um todo. O processo *ChurnSense* oferece a possibilidade do emprego individualizado destes algoritmos, bem como do emprego de maneira combinada, possibilitado diferentes situações de análise.

Para identificar perfis, o passo Análise assume que é suficiente empregar apenas duas métricas, a medida do RSSI e a quantidade de dispositivos conectados em determinada CPE. A hipótese assumida é que um cliente que esteja constantemente com um número elevado de dispositivos conectados e com o RSSI de seus dispositivos ruim (menor do que -65dBm) em sua CPE, estará mais exposto a uma conectividade instável, o que geralmente ocasiona insatisfação no cliente, podendo levar ao cancelamento do plano e ampliação do índice de *churn*.

O *ChurnSense* utiliza o algoritmo K-Means para segmentar os clientes do ISP em diferentes grupos. O K-Means tem como objetivo centralizar ao máximo os centroides com base na distância entre os pontos e o centro do agrupamento (Nandapala e Jayasena, 2020). Neste sentido, no *ChurnSense*, o processo que utiliza o K-Means tem os passos:

i) Definição do número de centroides ou *clusters*, representado por *num_clusters*;

ii) Treinamento e atualização dos centroides em *K* etapas, em que *K* faz referência o número de etapas de treinamento e atualização. Já o uso de variáveis utilizadas nesse momento compreendem o RSSI dos dispositivos e a quantidade de dispositivos conectados no roteador *QD*;

iii) Classificação, após o processo de treinamento, onde os dados estão agrupados em *clusters* e pertencem a um centroide (perfil) específico.

No passo *i*, o K-Means requer que o número de *clusters* seja definido a partir dos dados não rotulados, ou seja, *num_clusters* necessita ser determinado previamente.

No passo *ii*, o treinamento e a atualização dos centroides ocorrem através do cálculo das distâncias entre os pontos e os centroides. Esse processo matemático envolve a utilização da equação da distância euclidiana, representada na Eq. (1).

$$D = \sqrt{(x2 - x1)^2 + (y2 - y1)^2} \tag{1}$$

Nessa equação, as variáveis $(x1, y1)$ representam as coordenadas do ponto, enquanto que os termos $(x2, y2)$ representam as coordenadas do centroide. Considerando o conjunto de clientes do ISP, definido por C, onde $C = \{c_1, ..., c_n\}$, e dado um cliente c_i, onde $1 <= i >= n$, que possui um conjunto de dispositivos D, onde $D = \{d_1, ..., d_m\}$, que se conectam a um roteador r que fornece serviços ao cliente, uma distância D pode ser estabelecida entre cada dispositivo e seu respectivo roteador. Deste modo, um dispositivo d_i, onde $1 <= i >= m$, ao se conectar ao roteador, passa a possuir valores de RSSI $r.d_i.RSSI$ e, por sua vez, o roteador passa a observar a quantidade de dispositivos $r.QD$ em sua proximidade. Tais métricas são exploradas pelo *ChurnSense* para treinar a identificação de perfis de clientes.

Para exemplificar o processo de treinamento e atualização dos centroides, tomamos como referência os valores de *RSSI* e *QD* associados ao dispositivo d_i, ou seja, $r.d_i.RSSI > -65$ dBm e $r.QD \leq 10$. Após a conclusão das iterações do algoritmo K-Means, o cálculo da distância D (Eq. (1)) permite inferir que, nesse cenário específico, o cliente c_1 demonstra indícios sugerindo um potencial *churn*, devido ao sinal das conexões serem muito ruins e possuir um relevante número de dispositivos conectados em R.

No passo *iii*, ao final do processo de treinamento, os dados estarão agrupados e pertencerão a centroides específicos. Por exemplo, um cliente c_2 com métricas como $r.d_i.RSSI = -70$ dBm e $r.QD \leq 10$ também poderá apresentar indícios de propensão ao *churn*. A definição do perfil de cada resultado consiste em uma combinação das variáveis *RSSI* e *QD*. O tamanho do conjunto de dados, o número de etapas de treinamento K, o número de *clusters* *num_clusters* e a localização dos centroides são os principais fatores que influenciam nos resultados finais do algoritmo de K-Means. A escolha adequada dessas variáveis passa a ser crucial para obter agrupamentos significativos e úteis dos clientes do ISP.

No *ChurnSense*, além do *K-Means*, para encontrar agrupamentos em um conjunto de dados pode ser utilizada a Rede Neural SOM, que permite incrementar a compreensão e visualização das características dos dados. O seu uso tem como objetivo definir os grupos ou *clusters* de clientes do ISP com base em séries temporais e aprendizagem competitiva.

O funcionamento da Rede Neural SOM compreende três passos:

i) Inicialização e Organização, onde cada vetor recebe um peso com base na dimensionalidade do espaço;

ii) Competição e Ajuste, onde encontra-se o neurônio "vencedor"depois de todos os ajustes entre eles;

iii) Resultados e *Insights*, onde os resultados podem ser vistos e analisados, oferecendo capacidade de identificar

relações com os problemas e soluções.

No passo *i*, a Rede Neural SOM é aplicada na segmentação de clientes com base em diferentes características, assim como o K-Means faz com centroides. Porém, ao invés de usar centroides, a Rede Neural SOM cria um mapa auto-organizável, distribuindo neurônios em uma grade. Cada neurônio tem pesos que se ajustam aos dados de entrada. Durante o treinamento, neurônios similares competem, o vencedor se ativa e ajusta seus pesos, afetando vizinhos.

No passo *ii*, opera-se com duas camadas principais: a camada de entrada e a camada SOM. Dentro da camada SOM, os neurônios estão organizados em uma grade. Cada um desses neurônios faz referência a um cliente, por exemplo c_1, e está associado ao dispositivo, por exemplo d_1. Cada dispositivo tem a métrica chamada *QD*, que representa quantos dispositivos estão conectados ao mesmo roteador *R*. No início, os pesos de cada neurônio, incluindo o do cliente c_1 e o dispositivo d_1, são ajustados aleatoriamente. O processo de treinamento acontece em iterações. Durante cada iteração, um dado de entrada, que nesse caso inclui a métrica *QD* do dispositivo d_1, apresenta-se à rede. A partir dessa métrica, a rede neural seleciona um neurônio "vencedor", com base na similaridade dos pesos. O neurônio vencedor, juntamente com os neurônios vizinhos na grade, recebe ajustes para ficar mais semelhante ao dado de entrada. Essa etapa se repete para diferentes dados de entrada, como a *QD* de diferentes dispositivos, gradualmente ajustando os neurônios. Assim como anteriormente, a eficácia desse processo depende da organização dos neurônios, da quantidade de neurônios na rede e da influência da métrica *QD* para identificar padrões de agrupamento que possam indicar a probabilidade de *churn*.

No passo *iii*, para o contexto do *ChurnSense*, a aplicação da Rede Neural SOM permite agrupar clientes com base em padrões de uso, localização ou outras informações relevantes. Usando o mesmo exemplo de cliente c_1, por ele conter um alto número de dispositivos conectados simultaneamente ($QD \geq 7$) e vários picos de conexão, pode-se concluir que segundo a analise do SOM o cliente c_1 está propenso ao *churn*.

5 Estudo de Caso em um ISP Regional

Esta seção mostra o estudo de caso do processo *ChurnSense* empregando dados reais de um ISP de escala regional. A Seção 5.1 descreve o passo da Coleta dos Dados e Pré-Processamento de Dados, bem como ferramentas utilizadas. A Seção 5.2 apresenta o estudo de caso do passo de Análise. Por fim, a Seção 5.3 apresenta os resultados obtidos por meio de todo o processo, descrevendo detalhadamente cada perfil obtido, apontando os mais propensos ao *churn*.

5.1 Coleta de Dados e Pré-Processamento

Esta seção apresenta o estudo de caso dos passos Coleta de Dados e Pré-Processamento do processo *ChurnSense*. A Coleta de Dados foi implantada utilizando como base o servidor ACS de código aberto GenieACS (GenieACS, 2023), pois proporciona uma fácil configuração com códigos escritos em *JavaScript* (Afek et al., 2020). Ele também proporciona segurança na transmissão de pacotes, pois possibilita a implementação do protocolo SSL. Porém, cabe salientar que a aplicação do GenieACS requer uma configuração adicional juntamente com o uso de certificados digitais e aplicações de boas práticas (Hils e Böhme, 2020).

Todos os dados coletados no estudo de caso foram de roteadores da marca Huawei, variando entre os modelos, WS5200-21 (V2), WS-5200-40 (V3) e WS-7001-40 (AX2). Ao todo foram utilizados 9500 roteadores distribuídos entre as cidades que o ISP atende, localizadas na área Norte do estado do Rio Grande do Sul. A Coleta foi executada em um período de 15 dias (de 13 de fevereiro a 28 de fevereiro de 2023) e resultou em 870 conjuntos de dados com intervalo de 20 minutos entre conjuntos de dados.

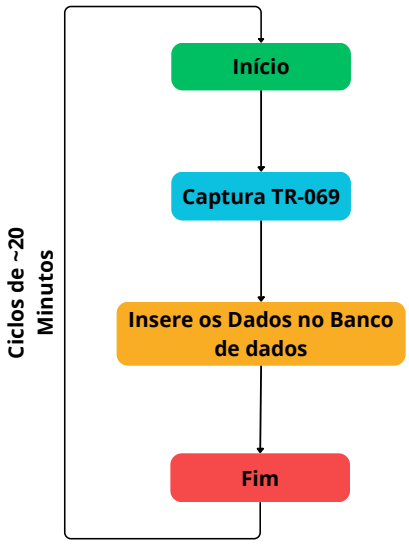

Figura 5: Diagrama Geral da Coleta de Dados

A Fig. 5 apresenta o fluxograma do funcionamento da Coleta dos Dados, dividido em duas etapas. Na primeira etapa, foi desenvolvido um *script* que interage diretamente com o GenieACS e que recolhe todos os dados armazenados no ACS. Na segunda etapa, os dados capturados foram salvos no Banco que Dados de Coleta, modelado para cumprir os requisitos do passo Coleta (descritos na Seção 4.1). O passo de Coleta é cíclico e se repete a cada 20 minutos, ou seja, após a inserção dos dados no banco de dados o script aguarda 20 minutos e reinicia o processo. Como resultado as séries temporais dos dados de coleta são formadas e armazenadas no Banco de Dados de Coleta.

Foram utilizadas três máquinas virtuais para todo do processo, cada uma contendo, respectivamente, o GenieACS, o Banco de Dados de Coleta e o Banco de Dados de Histórico. Ambos bancos de dados foram implantados utilizando o PostgreSQL. O Banco de Dados da Coleta armazena todos os dados brutos e tem um espaço de armazenamento de 100GB e o Banco de Dados de Histórico armazena apenas os dados de interesse do passo Análise e tem uma capacidade de armazenamento de 35GB.

No passo Pré-Processamento foi utilizado o Banco de Dados de Histórico para guardar as informações filtradas e gerar Séries Temporais específicas para o passo de Análise.

Neste trabalho, os dados selecionados consistem no RSSI e na Quantidade de Dispositivos Conectados. No GenieACS, estes dados correspondem aos seguintes parâmetros:

- *InternetGatewayDevice . LANDevice . *. WLANConfiguration . *. AssociatedDevice . *. AssociatedDeviceRssi* e
- *InternetGatewayDevice . LANDevice . *. WLANConfiguration . *. TotalAssociations*

A limpeza de dados do passo de Pré-Processamento consiste na remoção e padronização de informações que fogem do padrão esperado para aplicação ao modelo.

Durante o estudo de caso dos passos Coleta de Dados e Pré-Processamento de Dados, houveram algumas lições apreendidas. Informações disponibilizadas na documentação do protocolo TR-069 são algumas vezes insuficientes, se considerado a tecnologia utilizada. Destaca-se a coleta do parâmetro do RSSI, a compatibilidade dos modelos de roteadores Huawei, e boas práticas de segurança do TR-069.

Referente ao RSSI, o parâmetro disponível da documentação do protocolo TR-069 não era encontrado na listagem de parâmetros dos roteadores, pois foi implementado pela Huawei de uma forma diferente. Nesse sentido, foram explorados fóruns e documentações de terceiros para encontrar o parâmetro que continha o valor do RSSI na CPE (*InternetGatewayDevice . LANDevice . *. WLANConfiguration . *. AssociatedDevice . *. AssociatedDeviceRssi*), o que acabou resolvendo o problema, já que, com o parâmetro exposto, foi possível coletar os valores de RSSI. Também é importante destacar que nos modelos V2 e V3, quando empregado a imagem de *firmware "10.0.5.7(C947)"*, o CPE não suporta totalmente o protocolo TR-069, pois há parâmetros faltantes, como o próprio RSSI.

Outro desafio a ser destacado e que teve uma investigação focada nele, foi a segurança do GenieACS. A solução de geranciamento TR-069 do GenieACS permite a implementação do protocolo HTTPS, porém não ocorre de forma nativa e precisa ser realizada manualmente. No estudo de caso foram seguidas boas práticas de segurança destacas por Hils e Böhme (2020).

Com os passos de Coleta de Dados e Pré-Processamento finalizados, os dados para o passo de Análise encontra-se preparados e disponíveis no Banco de Dados de Histórico.

5.2 Análise

Com foco no passo Análise, esta seção apresenta os métodos utilizados no processo *ChurnSense* para a obtenção dos resultados, ou seja, apresenta o emprego do Método de *Elbow*, do algoritmo K-Means e da Rede Neural SOM.

Com o Método de *Elbow* (método do cotovelo é possível estimar o número de *clusters* ideal para o algoritmo K-Means. Aplicando o método aos dados do Banco de Dados de Histórico e observando a métrica WCSS (*Within-Cluster Sum-of-Squares*) para diferentes valores de K (número de *clusters*), o resultado (vide Fig. 6) permite identificar que o número de *clusters* deve ficara entre 3 a 4 ("cotovelo" indicado na figura) e que indica o início da estabilização da variância dentro de cada *cluster*.

Na amostra para o treinamento do algoritmo K-Means

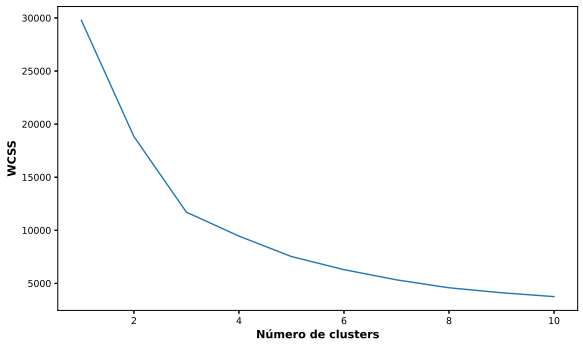

Figura 6: Método de *Elbow*

foi utilizado um total de 3019 roteadores de clientes do ISP em estudo, aos quais haviam 6541 dispositivos conectados. Foi optado por utilizar quatro *clusters*, pois os dados ficaram mais dispersos, facilitando a interpretação dos resultados.

O treinamento do K-Means levou em consideração os dados vindos do Pré-processamento, que neste estudo de caso correspondem ao número de dispositivos conectados e também ao RSSI de cada um dos dispositivos. Essas *features* foram usadas no eixo *X* do modelo, e como o algoritmo em questão segue uma abordagem não supervisionada, ou seja, onde não existem rótulos previamente definidos, não se fez presente a utilização do eixo *Y*, ou *target*. Por fim, devido a utilização de dados reais, as informações pessoais dos clientes em questão foram desprezadas.

Já a abordagem para o treinamento da Rede Neural SOM, teve uma amostra definida com 40 clientes do ISP e aproximadamente 350 dispositivos como *smartphones* e computadores destes mesmos clientes. O número de *clusters* seguiu o mesmo que no algoritmo K-Means, uma vez que a Rede Neural SOM não tem uma fórmula pré-definida para fazer esse cálculo. Porém, foi considerado somente a quantidade de dispositivos conectados e sua correspondente série temporal. Desta forma, as séries temporais possuem significativas diferenças, enriquecendo a análise.

5.3 Resultados

Esta seção descreve os principais resultados obtidos no processo *ChurnSense* pelos algoritmos K-Means e a Rede Neural SOM. O K-Means foi utilizado para classificar clientes em perfis e teve contribuição significativa para detectar clientes propensos ao *churn*. A Rede Neural SOM foi utilizada para avaliar o histórico de conexões simultâneas nos roteadores domésticos e teve contribuição significativa na identificação de picos em dias e horários específicos, os quais podem possuir relação com a qualidade de conectividade dos clientes.

A Fig. 8 mostra os resultados do algoritmo K-Means, revelando 4 grupos de perfis bem definidos, categorizados como: Muito Bom, Bom, Suficiente e Insuficiente. Os perfís são definidos como segue.

- Perfil **Muito bom**: apresenta a melhor qualidade de co-

nectividade, com um número baixo de dispositivos conectados e um RSSI muito bom. Portanto, assume-se que a chance de cancelamento é pequena neste perfil. Parâmetros de referência:

  - Menos de 10 dispositivos conectados.
  - RSSI até -45 (RSSI muito bom).
  - Total de 4778 dispositivos (20.32%)

- Perfil **Bom**: a conectividade é definida como boa, com um número baixo de dispositivos conectados e uma faixa de RSSI variada. A chance de cancelamento ainda é considerada pequena. Parâmetros de referência:

  - Dispositivos conectados abaixo de 8.
  - RSSI variado entre -45 a -65.
  - Total de 8240 dispositivos (35.07%)

- Perfil **Suficiente**: apresenta uma quantidade alta de dispositivos conectados, indicando maior demanda na rede, mas a qualidade da conectividade apresenta-se mediana ou ruim com um RSSI variando entre -25 a -95. Assume-se que a chance de cancelamento é média demandando monitoramento. Parâmetros de referência:

  - Maior número de dispositivos conectados (de 7 a 20+).
  - RSSI entre -25 a -95 (mediana/ruim).
  - Total de 3084 dispositivos (13.12%)

- Perfil **Insuficiente**: apresenta uma conexão de baixa qualidade, mesmo com um número reduzido de dispositivos conectados. O RSSI é considerado muito ruim, indicando problemas de conectividade significativos. Nesse perfil, a chance de cancelamento torna-se a alta e demanda ação rápida por parte do ISP. Parâmetros de referência:

  - Poucos dispositivos conectados (menos de 10).
  - RSSI entre -65 a -95 (muito ruim).
  - Total de 4866 dispositivos (20.61%)

Em resumo, quanto mais baixa a qualidade de atendimento, passa a ser maior a probabilidade de cancelamento do usuário. Perfis com qualidade de conexão melhor e menos dispositivos conectados têm uma menor probabilidade de cancelamento, enquanto perfis com qualidade de conexão pior e mais dispositivos conectados apresentam uma probabilidade mais alta de cancelamento.

O passo Análise delineou quatro perfis distintos de clientes com base na qualidade da conectividade, número de dispositivos conectados e força do sinal RSSI. O "Perfil Muito Bom"e o "Perfil Bom"destacam-se por oferecerem a melhor qualidade de conexão, com baixo risco de cancelamento. Esses perfis, caracterizados por um número reduzido de dispositivos conectados e RSSI favorável, indicam uma menor probabilidade de cancelamento.

Por outro lado, os perfis "Perfil Suficiente"e "Perfil Insuficiente"mostram sinais de maior demanda na rede e problemas significativos de conectividade. O "Perfil Insuficiente"destaca-se como o mais crítico, com uma probabilidade mais alta de cancelamento devido à baixa qualidade de conexão, mesmo com um número limitado de dispositivos conectados.

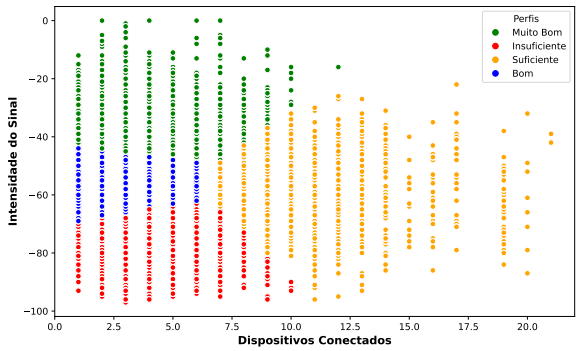

Figura 7: Perfis de Usuários sem escala

Diante desses resultados, é possível concluir que é imperativo que estratégias de retenção e melhoria de serviço se concentrem especialmente nos perfis de maior risco, como o "Perfil Insuficiente". A ação proativa por parte do ISP é crucial para mitigar os potenciais cancelamentos, enquanto aprimoramentos contínuos nos serviços podem fortalecer a fidelidade dos clientes nos perfis de menor risco.

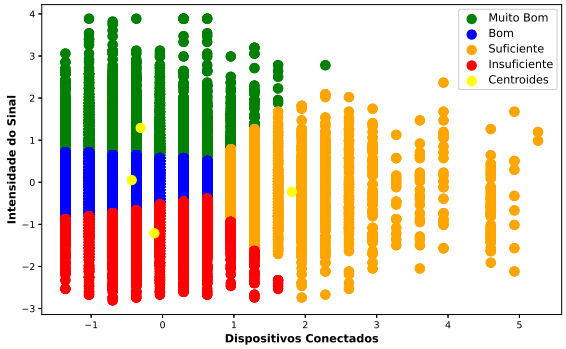

Figura 8: Perfis de Usuários em escala

O experimento de análise com a Rede Neural SOM, que considera a quantidade de dispositivos conectados em relação ao tempo, também apresenta 4 perfis de clientes. A série temporal que foi usada para esse modelo, foi coletada no período de 13 de fevereiro a 28 de fevereiro de 2023. Os resultados (ilustrados na Fig. 9) expressam o comportamento dos dispositivos em torno de um período de tempo e é possível analisar quando acontecem picos de conexão e quais perfis exigem uma atuação mais intensiva para resolução.

A partir dos resultados (Fig. 9) é possível analisar todos os horários (eixo X) juntamente com a quantidade de dispositivos conectados (eixo Y). Em horários entre as 10 horas e as 15 horas, observa-se uma conectividade maior (1 dispositivo a mais na média), o que pode indicar uma oscilação mais alta na qualidade da conectividade. Isso

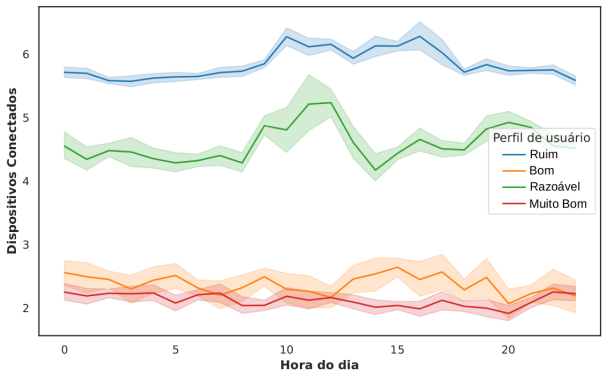

Figura 9: Comportamento dos dispositivos conectados ao longo de um dia

apresenta um possível problema, que, se não for tratado, pode levar o cliente ao *churn*.

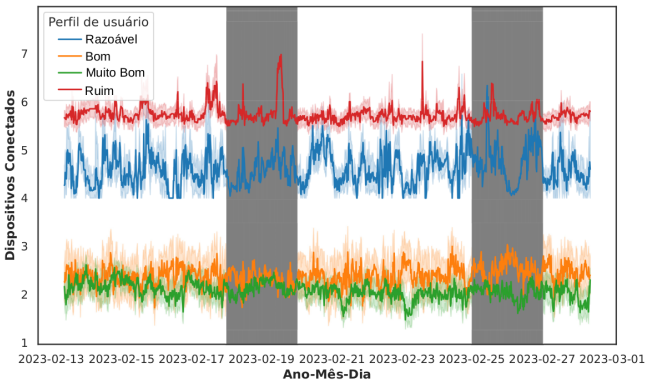

Figura 10: Quantidade de Dispositivos Conectados ao longo de 17 dias

Na Fig. 10, os finais de semana foram representados por barras verticais de cor cinza, onde indicam uma elevação de uma conexão na média de conexões nas CPEs dos clientes.

A partir destas análises (Fig. 9 e Fig. 10), os resultados podem ser interpretados como:

- Perfil **Razoável**:
  - Clientes com, em média, 5 dispositivos conectados.
  - Apresentam alta variação na quantidade de dispositivos (entre 2 a 3), aumentando ou diminuindo essa média no decorrer do período.
  - Devido ao número elevado de conexões simultâneas o roteador principal possivelmente irá se sobrecarregar, ocasionando uma conectividade instável de internet, o que pode levar o cliente ao *Churn*.

  Durante um dos finais de semana, no dia 25/02/2023, houve uma maior variação na média de conexões, o que contribuiu para a explicação dessa variabilidade. Essa variação pode ser resultado de diversos fatores, como atividades especiais, eventos ou promoções que levaram os clientes desse perfil a conectar mais dispositivos nesse período.
- Perfil **Bom**:
  - Clientes com, em média, 3 dispositivos conectados.
  - Apresentam variação na quantidade de conexões simultâneas de 1 (um) dispositivo.
  - Esses clientes apresentam uma boa qualidade devido ao baixo número de conexões simultâneas, tornando baixa a probabilidade do *Churn*.

  Permanecem na média durante os finais de semana, indicando que as atividades ou comportamentos desses clientes não são influenciados significativamente pelo período do final de semana, e a quantidade de dispositivos conectados permanece relativamente constante.
- Perfil **Muito Bom**:
  - Clientes com, em média, 2 dispositivos conectados.
  - Apresentam variação na quantidade de conexões simultâneas de 1 (um) dispositivo.
  - Clientes com a conectividade muito estável, dificilmente propensos ao *Churn*.

  Permanecem estáveis durante os finais de semana, indicando que as atividades ou comportamentos desses clientes não são fortemente influenciados pelo período do final de semana, e a quantidade de dispositivos conectados permanece relativamente constante.
- Perfil **Ruim**:
  - Clientes com, em média, 6 dispositivos conectados.
  - Apresentam variação baixa no geral (entre 0 e 1 dispositivos), com alguns picos de conexões (entre 1 e 2 dispositivos).
  - Devido a alta quantidade de dispositivos conectados, o ISP pode fazer a recomendação de um roteador secundário, esses clientes estão mais propensos ao *Churn*.

  Podemos perceber alguns picos de conexões em determinados momentos, indicando um aumento temporário na quantidade de dispositivos conectados. Além disso, nos finais de semana, há uma variação maior na quantidade média de dispositivos conectados em comparação com os dias úteis. Isso sugere que, durante os finais de semana, os clientes desse perfil podem ter um maior envolvimento com seus dispositivos, o que resulta em um aumento temporário na quantidade de dispositivos conectados. No entanto, essa variação ainda permanece dentro de uma faixa relativamente estável ao longo do tempo.

Cabe destacar que o experimento demonstra o potencial da Rede Neural SOM para analisar apenas o número de conexões dos clientes e alcançar classificações significativas que mapeiam clientes em perfis que permitem ações de mitigação de *churns*.

6 Discussão sobre Emprego dos Resultados

Esta seção apresenta uma discussão sobre como os resultados obtidos pelo processo proposto podem ser empregados por um ISP para elaborar planos de ação capazes de contri-

buir para a redução do *churn*. Em linhas gerais, a aplicação do processo revela perfis com indícios de um potencial cancelamento. Logo, os planos de ação podem ser empregados de modo a proporcionar uma melhoria da qualidade da experiência do cliente.

Elencamos duas principais estratégias que podem ser usadas, uma de cunho mais técnico e outra voltada para promover uma conscientização dos clientes. Considerando a perspectiva técnica, o ISP pode mobilizar uma equipe externa para agendar uma visita na residência do cliente com o objetivo de reposicionar o ponto de acesso dentro do espaço físico. Seguindo esta premissa, a presença de um profissional da área técnica assegura que a melhor área de cobertura foi escolhida para o cliente.

A segunda medida pode ser realizada via uma campanha de conscientização dos clientes. Neste caso, o ISP pode divulgar dicas nas suas redes sociais sobre as principais situações que degradam o desempenho do RSSI dos dispositivos nas redes domésticas e comerciais. Ao seguir esta abordagem, os próprios clientes podem contribuir para a melhoria da sua conectividade e passarem a não ser mais classificados em risco de cancelamento.

Estas abordagens apresentam pontos positivos e negativos, podendo inclusive serem empregadas simultaneamente. Ao mobilizar os técnicos para visitar os clientes, o ISP acaba arcando com os custos do deslocamento, mas em contrapartida, existe um profissional da área que assegura a qualidade do RSSI. Com uma campanha de conscientização, esses custos são menores, pois não necessariamente envolvem o deslocamento, todavia, podem existir clientes que não tenham condições de realizar os procedimentos técnicos para reposicionar o ponto de acesso de maneira apropriada. Logo, tais abordagens podem ser seguidas de maneira simultânea e gradativa pelo ISP, de forma a usufruir dos benefícios de cada estratégia.

7 Conclusões

Este trabalho apresentou o *ChurnSense*, um processo para auxiliar na identificação de perfis de clientes de ISPs propensos ao *churn* aplicando três passos: a Coleta, o Pré-Processamento e a Análise. Por meio destes passos, o *ChurnSense* permite reunir e tratar dados do protocolo TR-069 empregando algoritmos de ML como o K-Means e a rede neural SOM. Finalmente, o perfil de clientes com indícios de *churn* é revelado, permitindo que os ISPs atuem na sua mitigação. Um estudo de caso foi conduzido com dados reais obtidos de um ISP de escala regional. A partir de um servidor ACS de código aberto GenieACS foram coletados dados de 9500 roteadores Huawei utilizados pelo ISP. Os resultados obtidos mostram que com base apenas na análise de RSSI e número de dispositivos conectados, ambos obtidos via protocolo TR-069, é possível identificar grupos de clientes propensos ao *churn*. A compreensão do comportamento dos clientes fornece *insights* valiosos para estratégias de retenção e melhorias no atendimento e este trabalho mostra que isso pode ser feito com o emprego de dados do protocolo TR-069. Dentre as limitações deste trabalho, cabe ressaltar que a proposta depende do protocolo TR-069, o qual atualmente é suportado por poucos fabricantes e modelos de roteadores atualmente. Como trabalhos futuros, espera-se testar o processo *ChurnSense* em outros ISPs de escala regional a fim de obter resultados de empresas com diferentes características.

Agradecimentos

Ao Provedor de Serviços de Internet Tchê Turbo pelo financiamento desta pesquisa por meio da Fundação de Desenvolvimento da Pesquisa - FUNDEP.

Referências

Afek, Y., Bremler-Barr, A., Hay, D., Goldschmidt, R., Shafir, L., Avam, G. e Shalev, A. (2020). Nfv-based iot security for home networks using mud, *NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium*, IEEE, pp. 1–9. https://doi.org/10.1109/NOMS47738.2020.9110329.

Bachan, L. e Gaber, T. (2021). Predicting customer churn in the internet service provider industry of developing nations: A single, explanatory case study of trinidad and tobago, Vol. 1339, Springer Science and Business Media Deutschland GmbH, pp. 835–844. https://doi.org/10.1007/978-3-030-69717-4_77.

Bajpai, D. e He, L. (2020). Evaluating KNN performance on wesad dataset, *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, pp. 60–62. https://doi.org/10.1109/CICN49253.2020.9242568.

Basicevic, I. (2023). An analysis of the tr069 (cwmp) protocol, *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, IEEE, pp. 460–465. https://doi.org/10.23919/MIPRO57284.2023.10159841.

Campanile, L., Forgione, F., Marulli, F., Palmiero, G. e Sanghez, C. (2021). Dataset anonimyzation for machine learning: An isp case study, *in* O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blečić, D. Taniar, B. O. Apduhan, A. M. A. Rocha, E. Tarantino e C. M. Torre (eds), *Computational Science and Its Applications – ICCSA 2021*, Springer International Publishing, Cham, pp. 589–597. https://doi.org/10.1007/978-3-030-86960-1_42.

Dai, B., Cao, Y., Wu, Z., Dai, Z., Yao, R. e Xu, Y. (2021). Routing optimization meets machine intelligence: A perspective for the future network, *Neurocomputing* **459**: 44–58. https://doi.org/10.1016/j.neucom.2021.06.093.

Forum, B. (2020). Tr-069 cpe wan management protocol - version: 1.4, *Technical report*, Broadband Forum. Disponpivel em https://www.broadband-forum.org/pdfs/tr-069-1-6-1.pdf.

GenieACS (2023). Fast, lightweight tr-069 acs. Disponpivel em https://genieacs.com/.

Hils, M. e Böhme, R. (2020). Watching the weak link into your home: An inspection and monitoring toolkit for tr-069, *Applied Cryptography and Network Security: 18th International Conference, ACNS 2020, Rome, Italy, October 19–22, 2020, Proceedings, Part II 18*, Springer, pp. 233–253. https://doi.org/10.1007/978-3-030-57878-7_12.

Huang, K. S. (2019). An enhanced tr-069 firmware upgrade method of wi-fi mesh system, *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, IEEE, pp. 655–659. https://doi.org/10.1109/CCOMS.2019.8821760.

Ikhsan, R. B., Mohammed, G., Putriana, I., Sriwidadi, T., Aries e Wahono, J. W. (2022). Customer loyalty based on internet service providers-service quality, *2022 6th International Conference on Informatics and Computational Sciences (ICICoS)*, pp. 18–23. http://dx.doi.org/10.1109/ICICoS56336.2022.9930615.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B. e Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, *Information Sciences* **622**: 178–210. https://doi.org/10.1016/j.ins.2022.11.139.

Jain, H., Khunteta, A. e Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost, Vol. 167, Elsevier B.V., pp. 101–112. https://doi.org/10.1016/j.procs.2020.03.187.

Li, X. e Zhu, D. (2018). An adaptive som neural network method for distributed formation control of a group of auvs, *IEEE Transactions on Industrial Electronics* **65**(10): 8260–8270. https://doi.org/10.1109/TIE.2018.2807368.

Liao, C.-F. e Wang, L.-H. (2022). A blockchain-driven elastic firmware deployment platform for tr-069 compatible appliances, *2022 International Conference on Platform Technology and Service (PlatCon)*, pp. 1–6. https://doi.org/10.1109/PlatCon55845.2022.9932077.

Liao, J., Jantan, A., Ruan, Y. e Zhou, C. (2022). Multi-behavior rfm model based on improved som neural network algorithm for customer segmentation, *IEEE Access* **10**: 122501–122512. https://doi.org/10.1109/ACCESS.2022.3223361.

Lygerou, I., Srinivasa, S., Vasilomanolakis, E., Stergiopoulos, G. e Gritzalis, D. (2022). A decentralized honeypot for iot protocols based on android devices, *International Journal of Information Security* **21**(6): 1211–1222. https://doi.org/10.1007/s10207-022-00605-7.

Mingoti, S. A. e Lima, J. O. (2006). Comparing SOM neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms, *European Journal of Operational Research* **174**(3): 1742–1759. https://doi.org/10.1016/j.ejor.2005.03.039.

Mostacero-Agama, L. e Shiguihara, P. (2022). Analysis of internet service latency and its impact on internet of things (iot) applications, *2022 IEEE Engineering International Research Conference (EIRCON)*, pp. 1–4. https://doi.org/10.1109/EIRCON56026.2022.9934102.

N, T. R. e Gupta, R. (2020). A survey on machine learning approaches and its techniques:, *2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS)*, pp. 1–6. https://doi.org/10.1109/SCEECS48394.2020.190.

Nandapala, E. e Jayasena, K. (2020). The practical approach in customers segmentation by using the k-means algorithm, *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pp. 344–349. https://doi.org/10.1109/ICIIS51140.2020.9342639.

Otchere, D. A., Ganat, T. O. A., Gholami, R. e Ridha, S. (2021). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models, *Journal of Petroleum Science and Engineering* **200**: 108182. https://doi.org/10.1016/j.petrol.2020.108182.

Pebrianti, D., Istinabiyah, D. D., Bayuaji, L. e Rusdah (2022). Hybrid method for churn prediction model in the case of telecommunication companies, *2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pp. 161–166. https://doi.org/10.23919/EECSI56542.2022.9946535.

Peddarapu, R. K., Ameena, S., Yashaswini, S., Shreshta, N. e PurnaSahithi, M. (2022). Customer churn prediction using machine learning, *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pp. 1035–1040. https://doi.org/10.1109/ICECA55336.2022.10009093.

Prasetyo, Y. T., Ong, A. K. S., Nadlifatin, R., Ong, B. A. S., Adiguna, L. P., Asallie, I., Tanto, H. e Arden, K. (2022). Determining factors affecting customer loyalty to internet service provider during the covid-19 pandemic: A structural equation modeling approach, *2022 Asia-Pacific Computer Technologies Conference (APCT)*, pp. 21–31. https://doi.org/10.1109/APCT55107.2022.00009.

Rahman, M. D., Alam, M. D. e Hosen, M. D. (2022). To predict customer churn by using different algorithms, Institute of Electrical and Electronics Engineers Inc., pp. 601–604. https://doi.org/10.1109/DASA54658.2022.9765155.

Sambhwani, S., Bharadwaj, A., Drewes, C., Hamidouche, K., Naguib, A., Nickisch, D., Roessel, S., Sauer, M., Schoinas, Y., Tabet, T., Vallath, S. e Yu, Y.-T. (2022). Transitioning to 6g: Part 2-systems and network technology areas, *IEEE Wireless Communications* **29**(2): 6–8. https://doi.org/10.1109/MWC.2022.9801738.

Santos, G. H., Mendonça, G., de Souza, E., Leão, R. M., Menasché, D. S. et al. (2019). Análise nao supervisionada para inferência de qualidade de experiência de usuários residenciais, *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, SBC, pp. 958–971. https://doi.org/10.5753/sbrc.2019.7415.

Sarjonen, R. e Höyhtyä, M. (2023). Elbow estimation-based source enumeration method for lpi/lpd signals, *2023 Wireless Telecommunications Symposium (WTS)*, IEEE, pp. 1–6. https://doi.org/10.1109/WTS202356685.2023.10131679.

Servio, A., Oliveira, F., Dantas, J., Silva, D. e Clemente, D. (2023). Dependability issues on an internet service provider and availability study of autonomous systems, *2023 IEEE International Systems Conference (SysCon)*, pp. 1–7. https://doi.org/10.1109/SysCon53073.2023.10131183.

Streit, A. G., Leão, R. M., de Souza, E., Menasché, D. S. et al. (2019). Descobrindo perfis de tráfego de usuários: uma abordagem não supervisionada, *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, SBC, pp. 169–182. https://doi.org/10.5753/sbrc.2019.7358.

Tan, G., Zhuang, J., Zou, J. e Wan, L. (2022). Multi-type task allocation for multiple heterogeneous unmanned surface vehicles (usvs) based on the self-organizing map, *Applied Ocean Research* **126**: 103262. https://doi.org/10.1016/j.apor.2022.103262.

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K.-l. A., Elkhatib, Y., Hussain, A. e Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges, *IEEE Access* **7**: 65579–65615. https://doi.org/10.1109/ACCESS.2019.2916648.

Utamima, A., Aditya, F. I., Ashaari, F. W. G. e Nugraha, A. K. (2023). Applying machine learning algorithms for opinion mining on a digital internet service provider, *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, pp. 1–5. https://doi.org/10.1109/ISDFS58141.2023.10131886.

Vargas Maquilon, J. A. e Maruri Uriña, E. S. (2021). *Implementación de un servidor de autoconfiguración y monitoreo para un ISP ubicado en el Cantón Balzar.*, PhD thesis, Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas.

Wang, Y., Jia, M., Gao, N., Von Krannichfeldt, L., Sun, M. e Hug, G. (2022). Federated clustering for electricity consumption pattern extraction, *IEEE Transactions on Smart Grid* **13**(3): 2425–2439. https://doi.org/10.1109/TSG.2022.3146489.

Ximenes, D., Mendonça, G., Santos, G. H., de Souza, E., Leão, R. M., Menasché, D. S. et al. (2018). O problema de deteçao e localizaçao de eventos em séries temporais aplicado a redes de computadores, *Anais do XVII Workshop em Desempenho de Sistemas Computacionais e de Comunicação*, SBC. https://doi.org/10.5753/wperformance.2018.3323.

Yang, M.-S. e Hussain, I. (2023). Unsupervised multi-view k-means clustering algorithm, *IEEE Access* **11**: 13574–13593. https://doi.org/10.1109/ACCESS.2023.3243133.

Zhao, W., Queralta, J. P. e Westerlund, T. (2020). Sim-to-real transfer in deep reinforcement learning for robotics: a survey, *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 737–744. https://doi.org/10.1109/SSCI47803.2020.9308468.

Zhu, J. e Han, X. (2024). Big data clustering algorithm of power system user load characteristics based on k-means and som neural network, *Multimedia Tools and Applications* pp. 1–15. https://doi.org/10.1007/s11042-024-19156-1.