Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Sistemas de Informação

Predição de Carga de Autenticações em um Serviço Global de Acesso Sem Fio para Dimensionamento de Recursos

Samira Barroso dos Santos

JUIZ DE FORA
DEZEMBRO, 2023

Predição de Carga de Autenticações em um Serviço Global de Acesso Sem Fio para Dimensionamento de Recursos

Samira Barroso dos Santos

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Sistemas de Informação

Orientador: Edelberto Franco Silva

JUIZ DE FORA

DEZEMBRO, 2023

Predição de Carga de Autenticações em um Serviço Global de Acesso Sem Fio para Dimensionamento de Recursos

Samira Barroso dos Santos

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM SISTEMAS DE INFORMAÇÃO.

Aprovada por:

Edelberto Franco Silva
Doutor em Ciência da Computação

Alex Borges Vieira
Doutor em Ciência da Computação

André Luiz de Oliveira
Doutor em Ciência da Computação

JUIZ DE FORA
27 DE DEZEMBRO, 2023

Aos meus queridos pais, Maria Cristina e João Batista, que me apoiaram em todas as minhas decisões.

À minha graciosa irmã que sempre esteve ao meu lado.

Resumo

Com o aumento expressivo do número de usuários conectados em redes sem fio, a gestão desse volume de conexões se torna vital para otimizar economia e aprimorar a qualidade do serviço. Nesse contexto, este trabalho destaca a aplicação de técnicas e análises de inteligência computacional com o propósito de antecipar o crescimento da carga de usuários em um sistema global de autenticação sem fio. É importante observar que, apesar do crescente aumento na demanda por conectividade sem fio, as ferramentas atualmente disponíveis para o gerenciamento de redes Wi-Fi frequentemente apresentam deficiências. Isso inclui limitações na capacidade de prever com precisão o volume de conexões, planejar adequadamente a capacidade da rede e otimizar os recursos de autenticação. Tais deficiências muitas vezes resultam em sobrecarga dos servidores de autenticação e na degradação da qualidade do serviço. O objetivo principal desta pesquisa é identificar o momento adequado para reforçar a infraestrutura da rede e dos servidores de autenticação, com base na análise do volume de conexões ao longo de um período de tempo. Essa abordagem visa aprimorar a eficiência na gestão, no planejamento e na alocação de recursos computacionais relacionados aos servidores de autenticação. Como resultado desta investigação, temos que o modelo LSTM utilizado alcança valores de até 99% de precisão na predição do volume de usuários autenticados em um conjunto de dados reais. Aplicado na gerência de redes, resultados também comprovam que o modelo provê a utilização eficiente dos recursos computacionais relacionados ao enlace de conexão com a Internet.

Palavras-chave: Inteligência artificial, Aprendizado de Máquina, Predição de Carga de Usuários, Redes sem fio, Internet, LSTM

Abstract

With the significant increase in the number of users connected to wireless networks, managing this volume of connections becomes vital to optimize cost-efficiency and enhance service quality. In this context, this work highlights the application of computational intelligence techniques and analyses to anticipate the growth of user load in a global wireless authentication system. It is important to note that, despite the growing demand for wireless connectivity, the tools currently available for Wi-Fi network management often need to be improved. These include limitations in accurately predicting connection volumes, adequately planning network capacity and optimizing authentication resources. Such deficiencies often result in server authentication overload and service quality degradation. The main objective of this research is to identify the appropriate time to reinforce the network infrastructure and authentication servers based on the analysis of connection volumes over a period of time. This approach aims to enhance efficiency in managing, planning, and allocating computational resources related to authentication servers. As a result of this investigation, the LSTM model achieves up to 99% accuracy in predicting the volume of authenticated users in a real data set. Results also prove that the model efficiently uses computational resources related to the Internet connection link.

Keywords: Artificial Intelligence, Machine Learning, User Load Prediction, Wireless, Network

Agradecimentos

Aos meus pais que sempre se orgulharam de mim, me apoiaram e motivaram durante toda a minha trajetória.

À minha irmã, em quem eu me espelhava e sempre acreditou no meu potencial.

Ao meu orientador Edelberto, pela orientação, solicitude e apoio durante essa grande jornada da graduação.

Aos meus amigos, que me ajudaram e estiveram comigo durante esse momento.

"Many that live deserve death. And some that die deserve life. Can you give it to them? Then do not be too eager to deal out death in judgement. For even the very wise cannot see all ends."

J.R.R. Tolkien

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

DCC	Departamento de Ciência da Computução
UFJF	Universidade Federal de Juiz de Fora
QoE	*Quality of Experience*
QoS	*Quality of Service*
RNA	Rede Neural Artificial
IEEE	Instituto de Engenheiros Elétricos e Eletrônicos
DSL	*Digital Subscriber Line*
LAN	*Local Area Network*
BBS	Basic Service Set
DS	Distribution System
ESS	Extended Service Set
STA	Wireless Station
AP	Access Point
PSO	Otimização por Enxame de Partícula
SNMP	Simple Network Management Protocol
IP	Internet Protocol
IPFIX	P Flow Information eXport
MLP	Perceptron multicamadas
MSE	*Mean Squared Error*
RNN	Rede Neural Recorrente
MAPE	Média Percentual Absoluta do Erro
GRU	*Gated Recurrent Unit*
LSTM	*Long Short Term Memory*
SVR	Regressão de Vetor de Suporte
RNP	Rede Nacional de Ensino e Pesquisa
USP	Universidade de São Paulo

1 Introdução

1.1 Apresentação do Tema

A modelagem de padrões e a previsão de tráfego são questões importantes para as mais diversas aplicações em redes de comunicação sem fio, sejam elas locais, como as redes IEEE 802.11, ou redes metropolitanas, como nas redes 5G (Quinta Geração) (THANTHARATE et al., 2019). Diante do aumento significativo no número de usuários, aplicações e da carga de tráfego, os provedores de serviços sem fio devem atender a determinados requisitos. Dentre esses destaca-se a Qualidade de Serviço (QoS) das aplicações e a Qualidade de Experiência (QoE) dos usuários (GUO et al., 2021).

O núcleo de qualquer mecanismo de predição é monitorar o comportamento passado e presente de um sistema e, em seguida, estabelecer uma relação estatística entre o conjunto de entradas e o conjunto de saídas em uma determinada escala de tempo (estacionária) ou variante no tempo (não estacionária) (SUBRAHMANYAM; SATYANARAYANA, 2012). Destacamos que o tráfego de rede sem fio é considerado um sistema não linear e não estacionário (YOU; CHANDRA, 1999). A quantidade de usuários e os seus requisitos de rede são não lineares, devido às características de aplicação e o comportamento do usuário (TAE-EOG; LEE; SREENIVAS, 2006). Considerando isso, a utilização de Redes Neurais (RN) é um método eficiente para modelar, avaliar e prever o comportamento de sistemas não lineares e não estacionários (KIM et al., 2004).

Nesse contexto, este trabalho visa propor um modelo classificação/predição para prever a carga de usuários em uma rede sem fio, promovendo a possibilidade de uma melhora no gerenciamento dos recursos de alocação de largura de banda e no desempenho da rede.

1.2 Contextualização

As redes de acesso sem fio são os elementos chave para a implantação de cenários em que os usuários podem ter acesso. Seja esse acesso advindo de qualquer lugar e a qualquer hora, tanto para aplicações *Web* convencionais quanto àquelas emergentes, que exijam os requisitos de comunicação com a Internet. Neste contexto, muitos trabalhos têm sido realizados nos últimos anos, a fim de analisar, desenvolver e propor melhorias às tecnologias sem fios, incluindo a introdução de apoio à mobilidade e à QoS.

Nas últimas décadas, as redes sem fio vêm sofrendo um aumento significativo no número de usuários conectados devido à sua simples implementação técnica e baixos custos (MANZOOR et al., 2019). Essa notável expansão no crescimento também está relacionada a geração de novos aplicativos e modificações nos hábitos dos usuários.

Recentemente, com a restrição da locomoção de pessoas devido a pandemia do COVID-19, o comportamento das pessoas, os hábitos, e a forma como começaram a utilizar a Internet mudou significativamente. Como resultado, a caracterização da utilização da rede, como o tráfego das redes de acesso sem fio, foram afetadas de alguma forma, e junto vieram os desafios relacionados à velocidade e instabilidade das conexões.

Lidar com esse aumento da demanda por recursos de rede é de extrema importância em termos de economia energética e QoS. Como resposta a esses eventos, os administradores dessas redes precisam executar um bom gerenciamento e elaborar um planejamento apropriado para mapear todo o seu ambiente. Desta forma, é possível atender às demandas de uma conectividade de qualidade, com alta cobertura e alta velocidade.

1.3 Descrição do Problema

Em redes Wi-Fi - que são as redes locais sem fio no padrão IEEE 802.11 - há o problema associado à carga de usuários, que pode influenciar diretamente à QoS e à QoE associada. Isto é, quando muitas pessoas estão conectadas e usando ativamente a rede, a confiabilidade da transmissão em um meio sem fio tende a decair (*e.g.*, aumentando o atraso na transmissão de um quadro, incrementando a perda de pacotes, dentre outros efeitos negativos), principalmente em redes públicas. Usuários e aplicações podem experimentar

atrasos e interrupções devido a esses fatores. Esses efeitos são mais claramente observados quando há transmissão contínua de informação, *e.g.*, *streaming* de vídeos, jogos *online*, vídeo chamadas.

Além disso, do ponto de vista dos recursos da infraestrutura de transmissão, o mesmo problema pode ser observado. Os servidores e equipamentos podem ficar sobrecarregados com o aumento do número de usuários que utilizam o serviço. Essa sobrecarga pode resultar na rejeição de novos usuários, atrasos no processamento de solicitações, negação de serviço e outros problemas. Portanto, as organizações provedoras de redes de acesso precisam estar preparadas para lidar com uma demanda crescente.

Em razão disso, se faz necessário encontrar métodos capazes de identificar informações que auxiliarão no gerenciamento e tomadas de decisões. Isso pode envolver a avaliação de oportunidades para redução de custos, a determinação do momento adequado para aumentar a largura de banda ou o número de servidores de aplicação. Assim, o fortalecimento da rede com base na estimativa do número de usuários em uma rede Wi-Fi durante um determinado período de tempo se torna uma estratégia atraente, especialmente em um ambiente competitivo, no qual provedores de redes buscam otimizar seus gastos e aprimorar os serviços prestados.

1.4 Justificativa

Ao observar os desafios relacionados, tanto para a gestão, quanto à predição de carga de usuários em redes Wi-Fi, é possível identificar a necessidade de uma maior investigação com foco na descoberta e gerência com relação à robustez da rede. Questões cruciais surgem, como: “quando deve ser realizado o aumento da largura da banda? Ou quando deve-se alocar mais pontos de acesso (APs)?”, “Qual valor considerar como um limiar para determinar se a carga de usuários e acesso à rede é alto, ou baixo?”. Outro ponto relevante, que aponta para a economia de energia gasta por esses equipamentos é responder se, e quando, é possível desligar ou não seus equipamentos. Tais perguntas continuam em aberto no estado da arte do tema, como pode ser visto em (FRANK et al., 2021; RODRIGUES et al., 2022; APOSTOLO et al., 2022; PENG et al., 2016). Os trabalhos citados mostram um caminho para a utilização de técnica de inteligência computacional

e ressaltam a importância frente ao cenário de gerenciamento de redes.

Como benefícios, todos os usuários e administradores de redes envolvidos, quer sejam pessoas físicas ou jurídicas, podem se beneficiar com menos interrupções nas conexões, com ligações mais estáveis, aumento da produtividade e da qualidade experimentada, além de maior controle, gerência e possível redução de custos, especialmente quando APs são desligados por inatividade.

1.5 Objetivos

O presente trabalho tem o objetivo explorar a análise de séries temporais e avaliar a utilização de modelo de predição de carga, relacionados ao número de usuários conectados, em redes sem fio locais IEEE 802.11. Através de técnicas de Aprendizado de Máquina (*Machine Learning* - ML) (WANG; JIANG, 2018), serão analisados indicadores relevantes para "aprender com o passado e prever o futuro". Como destaque dos modelos utilizados, temos a adoção das redes neurais artificiais (RNA) como uma categoria de modelos computacionais promissora.

Este trabalho tem o objetivo de contribuir em três frentes, sendo elas, a caracterização, o planejamento e o gerenciamento de recursos relacionados a redes sem fio de larga escala. Para isso utiliza dados reais coletados da rede sem fio nacional da Rede Nacional de Ensino e Pesquisa.

1. **Caracterização:** compreender profundamente o comportamento e a utilização de serviços de autenticação em redes sem fio em nível nacional, utilizando técnicas de análise de séries temporais e redes neurais LSTM. Isso implica na identificação de padrões, tendências e comportamentos anômalos no tráfego da rede;

2. **Planejamento:** avaliar a viabilidade de redução de custos operacionais mantendo um desempenho satisfatório nos serviços de autenticação;

3. **Gerência:** administração dos recursos computacionais de maneira automatizada, com o desligamento e religamento desses servidores, alocação de maior ou menor quantidade de memória, banda, e outros recursos computacionais com base não só o número atual de autenticações mas a predição de carga futura;

4. **Aplicação da abordagem em um estudo de caso**.

1.6 Organização

Esta monografia está organizada em seis capítulos. O Capítulo 2 apresenta conceitos importantes para a compreensão da metodologia como: redes neurais artificiais, função de ativação e séries temporais. O Capítulo 3 apresenta os trabalhos relacionados utilizados para o desenvolvimento desta pesquisa. O Capítulo 4 apresenta a coleta e processamento dos dados, a metodologia que foi aplicada para a parte prática, apresentando a arquitetura desta proposta e detalhes do fluxo de trabalho. No Capítulo 5 são apresentados os resultados obtidos na experimentação. Por fim, o Capítulo 6 apresenta as conclusões.

2 Fundamentação Teórica

Para contextualização do problema a ser trabalhado neste documento, este capítulo apresenta a fundamentação teórica para compreender os conceitos que aqui serão apresentados. Primeiramente na Seção 2.1, serão apresentados informações sobre o surgimento do padrão IEEE 802.11 e a descrição dos componentes de sua arquitetura. Na sequência, a Seção 2.2 apresenta os conceitos de RNAs, detalhamento do funcionamento de um neurônio artificial, algumas funções de ativação, a arquitetura de uma rede e sobre redes LSTM. A Seção 2.3 traz os protocolos de gerência, descrevendo algumas formas possíveis de coleta de dados para o monitoramento e controle do desempenho de redes. A Seção 2.4 descreve um algoritmo otimizador de hiperparâmetros para modelos. Por fim, a Seção 2.5 estão as considerações finais para encerrar este capítulo.

2.1 Padrão IEEE 802.11

O padrão IEEE 802.11, atualmente conhecido como Wi-Fi, é indispensável na vida cotidiana. Conexões de Internet estão disponíveis em muuitos lugares através de pontos de acesso Wi-Fi ou pequenos roteadores *Digital Subscriber Line* (DSL)/Ethernet, permitindo que os dados sejam transferidos para dispositivos de um roteador/ponto de acesso.

No início da década de 90, o IEEE notou que a criação de um padrão de infraestrutura de comunicações sem fio era necessária para muitas empresas e instituições de pesquisa. Sendo assim, criaram um comitê de pesquisa denominado IEEE 802.11 que foi responsável pelo desenvolvimento de um padrão de rede sem fio que forneceria uma conexão confiável, rápida, barata e robusta (BERG, 2011).

Sua primeira versão surgiu em 1997, quando o IEEE publicou o padrão *Wireless Local Area Networks* (WLANs). A WLAN é um sistema que realiza a interconexão de diversos dispositivos, tanto móveis quanto fixos utilizando ondas de rádio de alta frequência como meio de transmissão, ao invés de cabos para conectar os dispositivos na *Local Area Network* (LAN). Os usuários que estão conectados por WLANs podem se movimentar

dentro da área de cobertura da rede.

Redes móveis podem ser classificadas de duas formas, infra-estruturada e independentes (ad-hoc). Na rede infra-estruturada, a transferência de dados acontece sempre entre uma estação e um ponto de acesso (do inglês - *Access Point*- AP) e nunca ocorre diretamente entre duas estações, ou seja, a rede infraestruturada é a que todos os clientes conectam no AP e não conversam diretamente entre si. Já em redes ad-hoc não requer infra-estrutura para funcionar. A comunicação é diretamente entre os entre as estações, se o destino não estiver ao alcance, requisita-se o serviço de outros *hosts* móveis vizinhos. Nenhum AP é necessário para controlar o acesso ao meio.

Neste trabalho o foco é a rede infraestruturada, que será mostrada na seção seguinte sobre a arquitetura.

A arquitetura é formada por múltiplos componentes que se integram para a construção do Wi-Fi (KUROSE; ROSS, 2013), esses componentes desta tecnologia serão ilustrados na Figura 2.1são descritos a seguir.

- **Conjunto de serviço básico (*Basic Service Set* - BSS)** – É o bloco de construção fundamental da arquitetura 802.11. Um BSS normalmente possui uma ou mais estações sem fio, que ficam sob o domínio direto de uma única função de coordenação. Esta determina o envio e recebimento de dados através do meio de transmissão sem fio utilizado pelas estações;

- **sistema de distribuição (*Distribution System* - DS)** – Componente lógico destinado a enviar quadros entre estações pertencentes a diferentes BSSs e uma rede local cabeada (LAN);

- **conjunto de serviço estendido (*Extended Service Set* - ESS)** – é uma interconexão entre vários BSSs permitindo que tenha uma área maior, já que as BSSs não tem um alcance tão grande quanto ao de uma ESS;

- **estação sem fio (*Wireless Station* - STA) -** Qualquer dispositivo que acesse o meio sem fio. Normalmente são desktops ou PCs (*Personal Computers*), onde possuem uma interface de comunicação sem fio; e

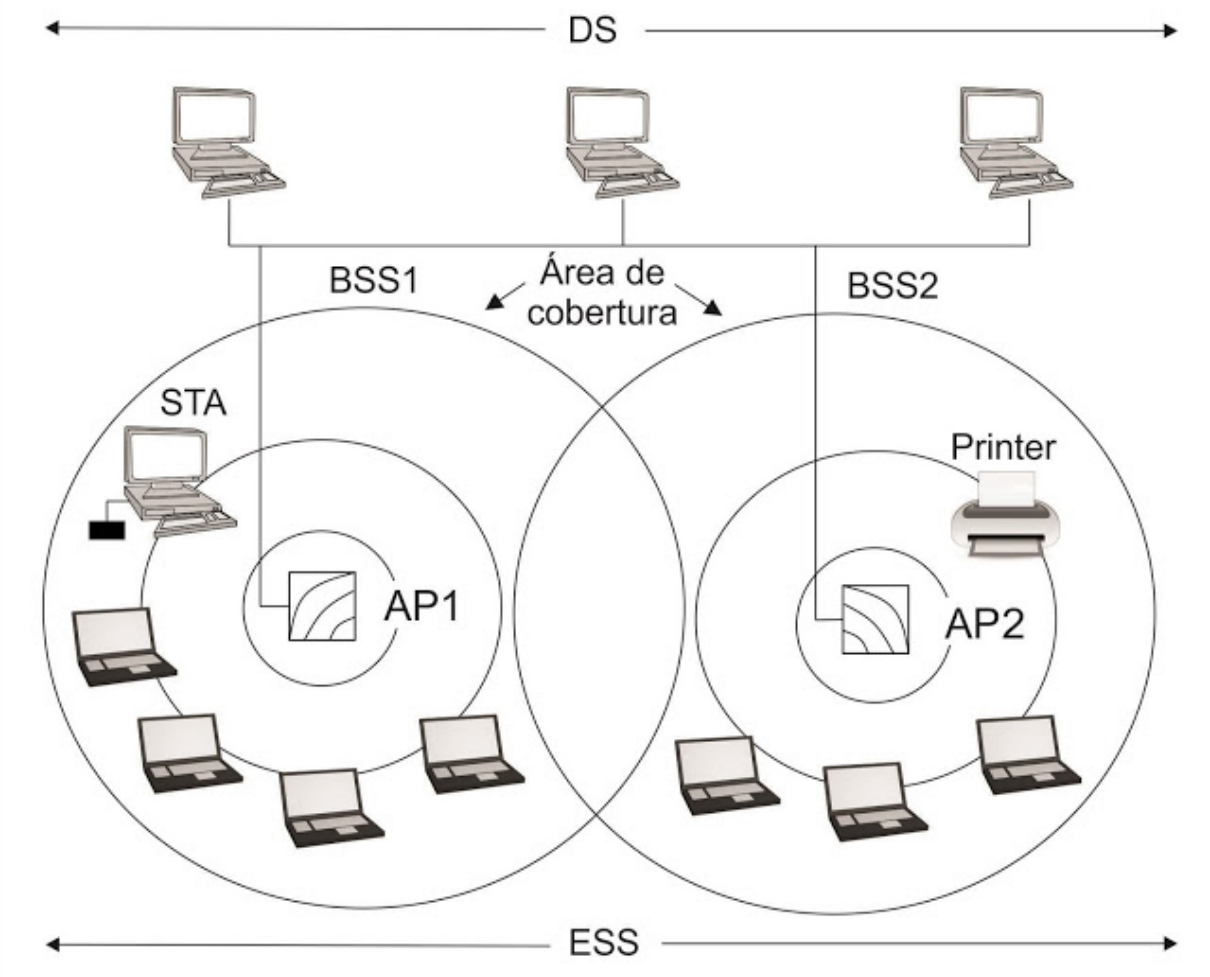

Figura 2.1: Arq. IEEE 802.11 (KUROSE; ROSS, 2013).

- **ponto de acesso (*Access Point* - AP)** – É uma estação STA que fornece conectividade entre vários STAs e entre STAs e DS.

O bloco de construção fundamental da arquitetura 802.11 é a célula, conhecida como BSS na linguagem 802.11. Um BSS normalmente contém uma ou mais estações sem fio e uma estação base central, conhecida como AP na terminologia 802.11. As estações, que podem ser fixas ou móveis, junto com a estação base central comunicam-se entre si usando o protocolo MAC sem fio IEEE 802.11. Vários APs podem ser conectados juntos, como por exemplo, usando uma Ethernet com fio ou outro canal sem fio, para formar o DS. O DS aparece para protocolos de nível superior, tal como o IP, como uma única rede 802, da mesma forma que uma rede Ethernet 802.3 com fio em ponte aparece como uma única rede 802 para os protocolos de camada superior (KUROSE; ROSS, 2013).

2.2 Rede Neural Artificial

A RNA é um sistema computacional que se inspira em redes neurais biológicas que constituem o cérebro animal e possui objetivo modelar o funcionamento do cérebro humano de forma digital, na tentativa de atingir a capacidade de processamento paralelo, não linear e altamente complexo do órgão animal (HAYKIN, 2001). Embora as RNAs sejam abstrações de uma rede neural biológica, a ideia das RNAs não é replicar o funcionamento dos sistemas biológicos, mas sim fazer uso do que se sabe sobre a funcionalidade das redes biológicas, servindo de modelo para o aprendizado e resoluções de problemas complexos.

A atratividade das RNAs vem das notáveis características de processamento de informações do sistema biológico, como não linearidade, alto paralelismo, robustez, tolerância a falhas e falhas, aprendizado, capacidade de lidar com informações imprecisas e difusas e sua capacidade de generalização (JAIN; MAO; MOHIUDDIN, 1996).

2.2.1 Neurônio Artificial

Uma RNA pode ser criada simulando uma rede de neurônios modelo em um computador. Um neurônio modelo é uma unidade de processamento de informação que é fundamental para a operação de uma RNA (HAYKIN, 2001).A Figura 2.2 ilustra o modelo de um neurônio, que forma a base para o projeto de RNA. Alguns elementos básicos do modelo neuronal são identificados:

- sinais de entrada $\{x_1, x_2, ..., x_m\}$: são os sinais externos normalizados para incrementar a eficiência computacional dos algoritmos de aprendizagem. São os dados que alimentam seu modelo preditivo. Um sinal x_j na entrada da sinapse j conectada ao neurônio k é multiplicado pelo peso sináptico w_{kj};

- pesos sinápticos $\{w_{k1}, w_{k2}, ..., w_{km}\}$: são valores para ponderar os sinais de cada entrada da rede. Esses valores são aprendidos durante o treinamento. O primeiro índice se refere ao neurônio em questão e o segundo se refere ao terminal de entrada da sinapse à qual o peso se refere;

- combinador linear $\{\sum\}$: agregar todos sinais de entrada que foram ponderados pelos respectivos pesos sinápticos a fim de produzir um potencial de ativação;

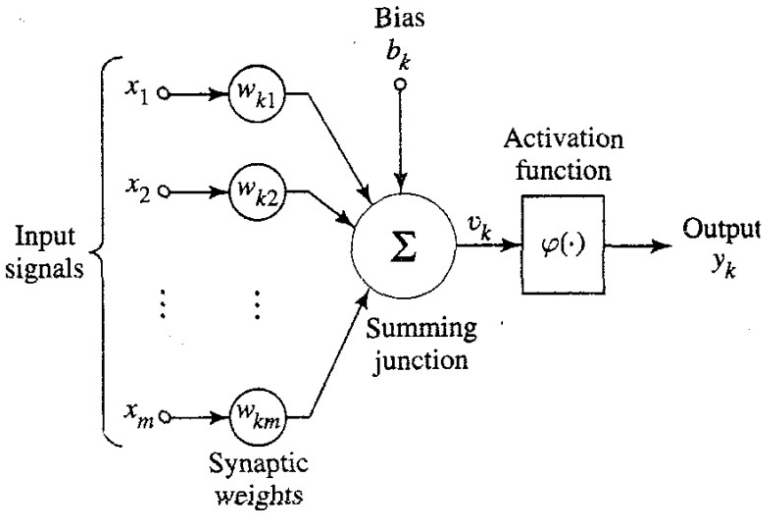

Figura 2.2: Modelo de um neurônio artificial (HAYKIN, 2001)

- limiar de ativação ou bias $\{b_k\}$: especifica qual será o patamar apropriado para que o resultado produzido pelo combinador linear possa gerar um valor de disparo de ativação;

- potencial de ativação $\{v_k\}$: é o resultado obtido pela diferença do valor produzido entre o combinador linear e o limiar de ativação. Se o valor for positivo, ou seja, se $v \geq 0$ então o neurônio produz um potencial excitatório; caso contrário, o potencial será inibitório;

- função de ativação $\{\varphi(.)\}$: seu objetivo é limitar a saída de um neurônio em um intervalo valores;

- sinal de saída $\{y_k\}$: é o valor final de saída podendo ser usado como entrada de outros neurônios que estão sequencialmente interligados.

Em termos matemáticos, a saída de um neurônio k pode ser descrita a partir da Equação 2.1.

$$y_k = \varphi(v_k) = \varphi(\sum_{j=1}^{m} x_j \times w_{kj} + b_k) \tag{2.1}$$

O neurônio processa as entradas de várias outras unidades ou fontes externas, pesa cada entrada e as soma. Em seguida, é aplicada uma função de ativação sobre

seu resultado. O resultado obtido é enviado para outros neurônios, caso exista mais de uma camada na rede, ou representará a saída da RNA, caso esteja localizado na última camada.

2.2.2 Função de Ativação

Em uma RNA, uma função de ativação é uma função matemática que determina se uma determinada entrada deve ser ativada ou não. Vários tipos de função de ativação podem ser usadas, elas definem a saída de um neurônio (HAYKIN, 2001).

Ao observar a Figura 2.2 novamente, nota-se que a função de ativação é alimentada pela soma de entradas e pesos combinados com o bias. Existem muitas funções de ativação diferentes que podem ser usadas, e a escolha da função de ativação pode ter um impacto significativo no desempenho da rede neural. As funções de ativação mais comuns usadas em uma rede neural são a função Sigmoid, a função Tanh e a função ReLU.

Algumas representações gráficas das funções de ativaçãosão ilustradas na Figura 2.3.

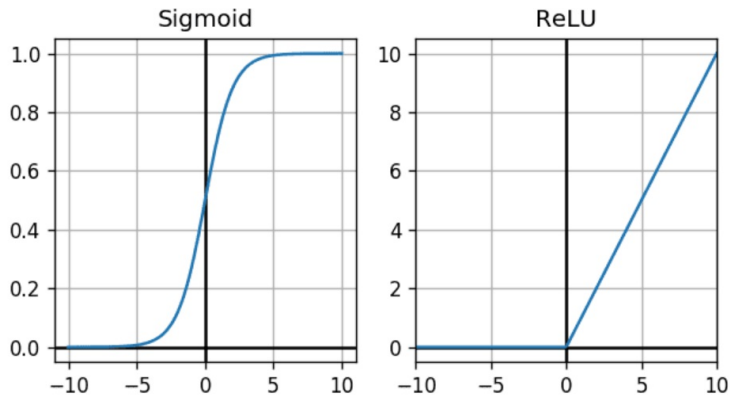

Figura 2.3: Representação gráfica da função Sigmoid e função ReLu

1. **Função Sigmoid**

   Segundo Haykin (2001), a função Sigmoid é definida como uma função estritamente crescente que exibe um balanceamento adquado entre comportamento linear e não-linear. Um exemplo dessa função é a função logística, definida pela Equação 2.2 e sua derivada na Equação 2.3,

$$\varphi(v) = \frac{1}{1 + e^{-av}} \tag{2.2}$$

$$\varphi'(v) = \frac{ae^{-av}}{(1 + e^{-av})^2} \tag{2.3}$$

onde a é o parâmetro de inclinação da função Sigmoid.

2. **Função ReLU**

   A *Rectified Linear Unit* (ReLU) é uma função não-linear e computacionalmente eficiente que retorna um valor positivo ou 0. Ela é amplamente utilizada e é a escolha padrão, pois produz melhores resultados. Sua representação matemática é definida através da Figura e da Equação 2.4 e sua derivada da Equação 2.5.

$$\varphi(v) = max(0, v) \tag{2.4}$$

$$\varphi^{'}(v) = \begin{cases} 1 & \text{se } v > 0 \\ 0 & \text{se } v \leq 0 \end{cases} \tag{2.5}$$

2.2.3 Arquitetura da rede

As redes neurais podem ser categorizadas em diversas tipologias, cada uma delas projetada para atender a propósitos específicos. Sua categorização é baseada em vários fatores, incluindo a profundidade, o número de camadas ocultas e as capacidades de entrada e saída de cada nó.

As arquiteturas mais usuais das RNAs são:

1. Redes Neurais Não Recorrentes

   As redes neurais não recorrentes, caracterizam-se pela inexistência de conexões entre neurônios da mesma camada, camadas anteriores ou camadas não subsequentes. Essa arquitetura segue uma única direção, em que os neurônios de uma camada são gerados a partir dos resultados das camadas anteriores.

Esses tipos de redes podem ser categorizados em dois principais arranjos: o de camada única e o de múltiplas camadas. No caso da camada única, observa-se que há apenas a presença da camada de saída dos neurônios, conforme ilustrado na Figura 2.4. Essa configuração é adequada para tarefas mais simples em que a relação entre entrada e saída é direta.

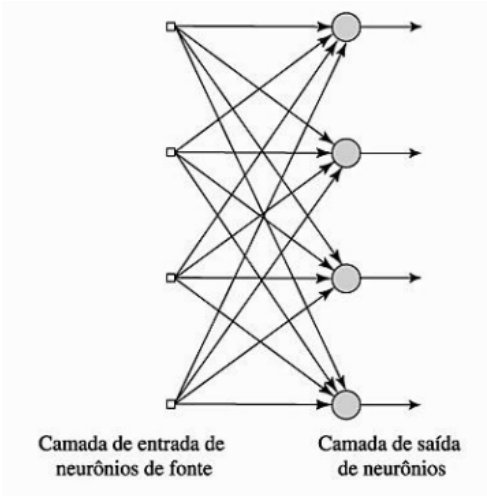

Figura 2.4: Exemplo de rede neural não recorrente de camada única (HAYKIN, 2001)

Por outro lado, as redes neurais não recorrentes com múltiplas camadas apresentam uma estrutura mais complexa, incorporando uma ou mais camadas ocultas entre os neurônios de entrada e saída, conforme exemplificado na Figura 2.5. Essas camadas ocultas têm o papel de aprender representações mais abstratas e complexas dos dados de entrada, permitindo que a rede realize tarefas mais sofisticadas e extrai características mais profundas. (HAYKIN, 2001)

2. Redes Neurais Recorrentes (*recurrent neural network* - RNNs)

   As redes neurais recorrentes são caracterizadas pela presença de realimentação entre neurônios da mesma camada ou entre camadas não subsequentes. Essa arquitetura, conforme ilustrado na Figura 2.6, introduz uma maior complexidade, resultando em uma melhoria na capacidade de aprendizagem. No entanto, é importante destacar que essa complexidade adicional pode afetar o desempenho da rede, uma vez que os valores de saída das camadas tornam-se dependentes não apenas das entradas atuais, mas também das saídas anteriores.

   A realimentação em redes neurais recorrentes possibilita que a informação seja re-

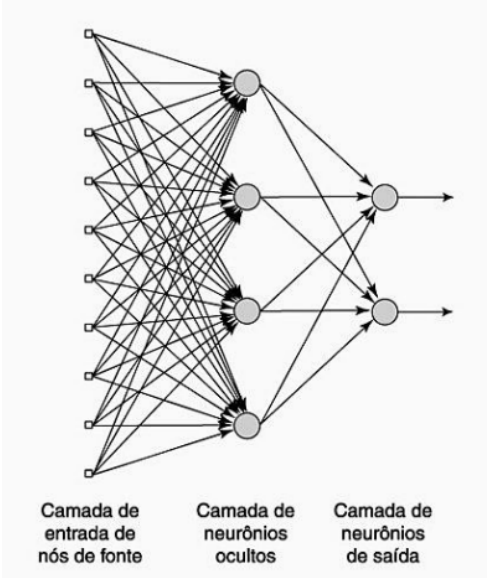

Figura 2.5: Exemplo de rede neural não recorrente de múltiplas camadas (HAYKIN, 2001)

troalimentada ao longo do tempo, permitindo que a rede mantenha uma espécie de memória interna. Isso torna essas redes particularmente eficazes em tarefas que envolvem sequências temporais, como séries temporais, processamento de linguagem natural e reconhecimento de fala.

A melhoria na capacidade de aprendizagem é resultado da capacidade da rede de capturar dependências temporais e padrões de longo prazo nas sequências de dados. No entanto, essa vantagem vem acompanhada de um aumento na complexidade computacional e na dificuldade de treinamento. O fenômeno conhecido como o problema do gradiente desvanecente pode surgir, dificultando o ajuste eficiente dos pesos durante o treinamento.

Portanto, embora as redes neurais recorrentes sejam poderosas para lidar com dados sequenciais e temporais, é necessário considerar cuidadosamente os compromissos entre complexidade e desempenho ao aplicá-las em diferentes contextos. (HAYKIN, 2001)

2.2.4 Long Short-Term Memory

Um desafio comum nas RNNs é a dificuldade em reter informações de longo prazo durante o treinamento. Esse problema motivou o desenvolvimento do modelo arquitetural para redes neurais recorrentes chamado de LSTM (*Long Short-Term Memory*), projetado

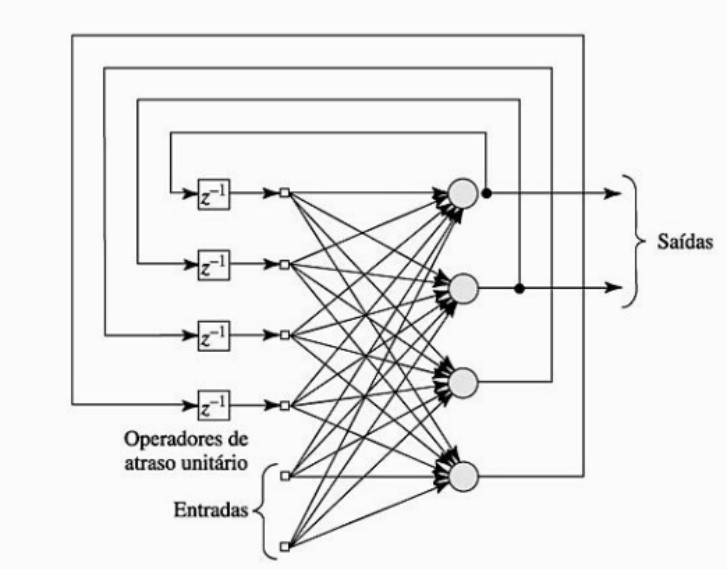

Figura 2.6: Exemplo de rede neural recorrente com neurônios ocultos (HAYKIN, 2001)

para superar essa limitação, permitindo a retenção efetiva de informações por períodos prolongados (HOCHREITER; SCHMIDHUBER, 1997). Isso se mostra particularmente vantajoso em tarefas relacionadas a sequências temporais, dada a tendência desses conjuntos de dados em serem extensos.

Cada modelo LSTM é composto por células LSTM repetidas, que desempenham as mesmas operações para cada elemento de dados fornecido a elas. Essas células possuem três portas fundamentais: a porta de esquecimento, a porta de entrada e a porta de saída. Essa arquitetura específica permite que as LSTMs superem as limitações de retenção de informações de curto prazo associadas às RNNs tradicionais, tornando-as especialmente eficazes em contextos onde a memória de longo prazo é crucial, como em problemas envolvendo sequências temporais extensas (SALAM et al., 2021).

O LSTM é composto por células de memória interconectadas que armazenam informações por um período de tempo arbitrário. Cada célula de memória é conectada à célula de memória anterior e à célula de memória seguinte por meio dos portões de entrada e saída. A célula de memória também está conectada a si mesma por meio do portão de esquecimento. Isso permite que as células de memória armazenem informações por um período de tempo arbitrário e aprendam dependências de longo prazo. Essas células são controladas por três portões: o portão de entrada, o portão de saída e o portão de esquecimento. Esses portões são conhecidos como *forget gate*, *input gate* e *output gate*

(Figura 2.7).

O *forget gate* controla a quantidade de informação antiga que é mantida na célula de memória e qual é descartada. Este portão recebe como entrada o parâmetro $x(t)$, que representa a entrada atual, e $h(t-1)$, que corresponde à saída do estado anterior. Esses valores são submetidos a uma função sigmoide, cujo resultado é representado por $f(t)$. Este último é um valor numérico situado entre zero e um, onde proximidade a zero indica esquecimento e proximidade a um indica armazenamento.

O *input gate* controla a quantidade de nova informação que entra na célula de memória. Mais uma vez, os parâmetros $x(t)$ e $h(t-1)$ são inicialmente submetidos a uma função sigmoide. Essa etapa é fundamental para determinar a relevância do valor na atualização, sendo que o resultado da função é representado por $i(t)$ e varia entre zero e um. Zero indica baixa importância, enquanto um indica alta importância. Simultaneamente, os parâmetros $x(t)$ e $h(t-1)$ também passam por uma função tangente hiperbólica, resultando em um novo vetor $\tilde{c}(t)$ com possíveis valores de atualização. Finalmente, realizamos a multiplicação das saídas das duas funções, com o propósito de que o resultado da sigmoide indique a importância da informação presente em $c(t)$, que é o resultado da tangente hiperbólica. Dessa forma, busca-se evitar a redundância e aprimorar a expressão original.

O *output gate* controla a quantidade de informação que é transmitida para a próxima célula de memória ou para a camada de saída. Inicialmente, os parâmetros $x(t)$ e $h(t-1)$ são novamente processados por uma função sigmoide, gerando $o(t)$. Esse resultado é então multiplicado pelo valor resultante da função tangente hiperbólica, que considera a soma das informações geradas pelos dois primeiros portões (representado por $c(t)$). O produto resultante é denotado como $h(t)$, representando o estado oculto. Ao mesmo tempo, o valor de $c(t)$ é atualizado para se tornar o novo conteúdo da célula. Em resumo, este portão é responsável por identificar quais parâmetros de entrada são mais relevantes para a saída. (HOCHREITER; SCHMIDHUBER, 1997).

As equações a seguir mostram os cálculos feitos em cada etapa, onde x_t é o vetor de entrada na etapa de tempo t, $h_t - 1$ é a saída da célula de memória na etapa de tempo anterior e $c_t - 1$ é a memória da célula anterior:

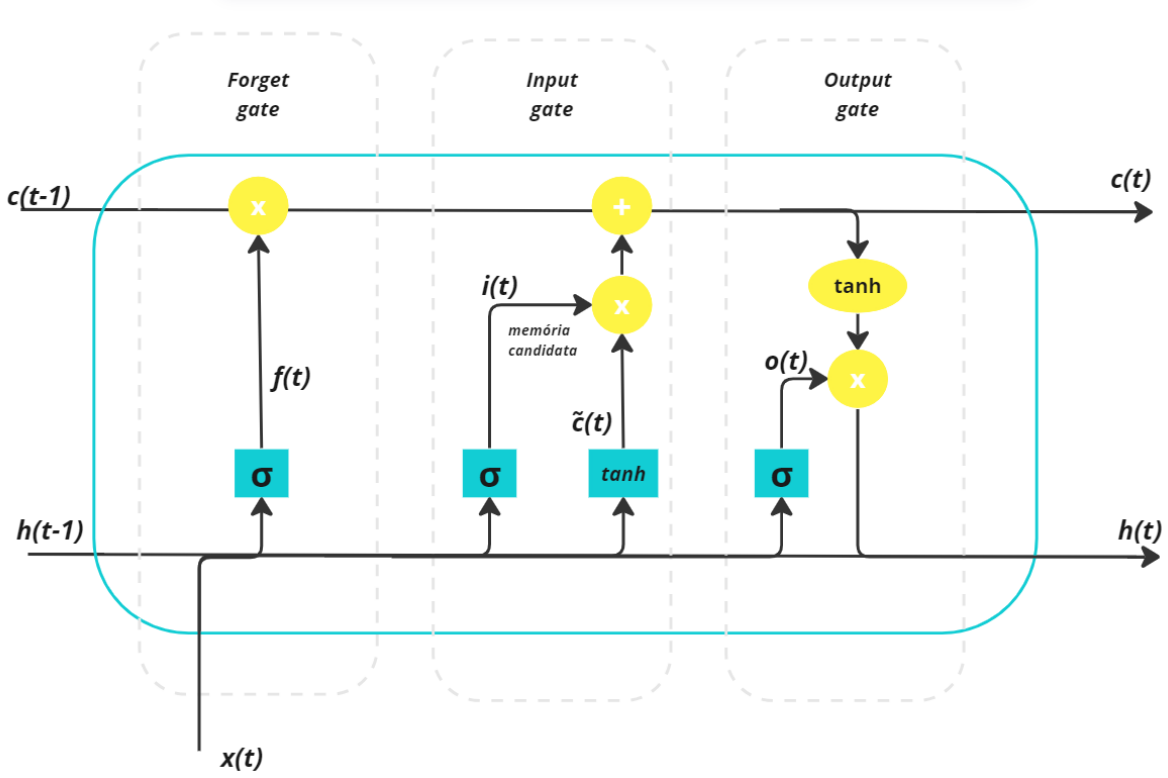

Figura 2.7: Arquitetura célula LSTM
Elaborado pela autora. (2023)

1. **Forget gate**

$$f_t = \sigma(W_f \cdot h_{t-1} + W_f \cdot x_t + b_f) \tag{2.6}$$

W_f é a matriz de peso de f, b_f é o bias de f e σ é a função sigmoide.

2. **Input gate**

$$i_t = \sigma(W_i \cdot h_{t-1} + W_i \cdot x_t + b_i) \tag{2.7}$$

W_i é a matriz de peso de i, b_i é o bias de i e σ é a função sigmoide.

$$\tilde{c}_t = \tanh(W_{\tilde{c}} \cdot h_{t-1} + W_{\tilde{c}} \cdot x_t + b_{\tilde{c}}) \tag{2.8}$$

$W_{\tilde{c}}$ é a matriz de peso de $\tilde{c}$, $b_{\tilde{c}}$ é o bias de $\tilde{c}$ e tanh é a função tangente hiperbólica.

3. **Atualização célula**

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \tag{2.9}$$

4. **Output gate**

$$o_t = \sigma(W_o \cdot h_{t-1} + W_o \cdot x_t + b_o) \tag{2.10}$$

W_o é a matriz de peso de o, b_o é o bias de o e σ é a função sigmoide.

$$h_t = o_t \cdot \tanh(c_t) \tag{2.11}$$

Até o atual momento, o LSTM foi aplicado com sucesso em tarefas como reconhecimento de escrita, tradução e reconhecimento de fala. Ao contrário da rede neural profunda comum (DNN) ou da rede neural convolucional (CNN) que possui apenas um caminho direto, o LSTM possui um loop que realimenta a saída da rede para a entrada e uma unidade de memória que mantém o estado da rede. A estrutura do loop e a unidade de memória permitem que as informações persistam entre etapas contínuas, fazendo com que a rede LSTM tenha a capacidade de processar dados de séries temporais (DING et al., 2020).

2.3 Protocolos de Gerência

O monitoramento de rede pode ser classificado em monitoramento ativo e passivo. As abordagens ativas, implementadas por ferramentas comuns, como o *ping* e *traceroute*, injetam tráfego em uma rede para realizar diferentes tipos de medições. Por outro lado, as abordagens passivas observam o tráfego existente à medida que ele passa por um ponto de medição, observando o tráfego do usuário. A abordagem de monitoramento passivo captura pacotes e geralmente fornece mais informações sobre o tráfego de rede, visto que pacotes completos podem ser capturados e analisados (HOFSTEDE et al., 2014)

2.3.1 *Simple Network Management Protocol* (SNMP)

De acordo com Medeiros et al. (2019), o SNMP é usado para coletar dados relacionados a alterações de rede ou para determinar o status de dispositivos conectados à rede. Cada dispositivo dentro da rede pode ser consultado em tempo real com SNMP, TCP e outros tipos de sondas para suas métricas de desempenho. Quando os limites de determinados

valores são excedidos, o software pode alertar os administradores do sistema sobre o problema, que podem acessar essas informações e, até mesmo, mudar o estado de alguma propriedade do elemento que está sendo monitorado, permitindo que eles analisem os dados e solucionem o problema.

O SNMP apresenta algumas limitações bem conhecidas, ele não é apropriado para o gerenciamento de redes muito grandes. Como ele é baseado em um mecanismo de sondagem *(polling)*, gerenciar muitos elementos de rede pode provocar atraso excessivo (MEDEIROS et al., 2019).

2.3.2 *NetFlow*

Segundo Medeiros et al. (2019), o NetFlow é um protocolo de rede desenvolvido pela Cisco para coletar informações de tráfego *Internet Protocol* (IP) e monitorar o fluxo da rede. Ele permite a coleta e agregação de informações sobre o tráfego de rede que entra ou sai de um elemento de rede que tem o protocolo habilitado. Os registros de informações coletados pelo NetFlow são enviados por mensagens do protocolo a um local centralizado na rede. Nesse local, os dados podem ser analisados por um administrador de rede que pode determinar, entre outros, a origem e o destino do tráfego, as classes de serviço de tráfego na rede e causas de possíveis congestionamento na rede (MEDEIROS et al., 2019).

2.3.3 *IP Flow Information eXport* (IPFIX)

O IPFIX é um padrão derivado do NetFlow para exportação de dados sobre os fluxos de rede. É um protocolo flexível que suporta campos de comprimento variável. Ele permite coletar informações como URL ou host e rastreia as ações de IP na rede. Para isso, o IPFIX coleta pacotes de dados de toda a rede, que são então organizados por um Exportador, que envia as informações compiladas para um Coletor. No IPFIX, os Exportadores podem transportar dados para vários Coletores, o que é conhecido como relacionamento muitos para muitos. Os exportadores enviam conjuntos de informações por meio de mensagens IPFIX, usando modelos especiais compostos por vários elementos (MEDEIROS et al., 2019).

2.3.4 *Logs*

No contexto de software, um *log* é o registro criado automaticamente e com registro de data e hora de ocorrências relevantes para um sistema específico. Programas e sistemas de software geram arquivos de *log*. Eles contêm detalhes sobre o aplicativo, dispositivo do usuário, hora, endereço IP e muito mais.

O gerenciamento de *logs* é um processo essencial da manutenção de um sistema saudável, pelo qual administradores observam continuamente os *logs* à medida que são gravados. O monitoramento de *log* é empregado para coletar, examinar e compreender os dados de desempenho da rede, monitorar e rastrear erros, para garantir que balanceadores de carga , roteadores e *firewalls*, por exemplo, estejam funcionando corretamente. Além disso, ajuda a criar comunicações seguras, auditar e corrigir problemas de rede.

2.4 Otimização de Hiperparâmetros com Optuna

A sintonia adequada dos hiperparâmetros é um aspecto fundamental na construção de modelos de aprendizado de máquina, diretamente influenciando o desempenho desses modelos. Nos últimos anos, a pesquisa dedicada à otimização de hiperparâmetros alcançou avanços significativos, resultando em diversas soluções inovadoras. A capacidade de ajustar esses parâmetros de forma precisa não apenas melhora a eficácia preditiva dos modelos, mas também fortalece sua adaptabilidade a diferentes cenários e conjuntos de dados. Esse domínio dinâmico desempenha um papel vital na busca contínua por modelos mais eficientes e generalizáveis, ressaltando a importância persistente da otimização de hiperparâmetros no cenário em constante evolução do aprendizado de máquina (PINHEIRO; BECKER, 2023).

Existem um grande número de ferramentas notáveis com objetivo de otimizar hiperparâmetros, no entanto, cada uma delas apresenta uma abordagem única em termos de usabilidade, podendo variar em flexibilidade conforme o contexto ou complexidade do modelo. Isso implica que a escalabilidade dessas aplicações pode ser comprometida (PINHEIRO; BECKER, 2023).

Diante desse cenário, surge o Optuna, um *framework* de código aberto destinado

à otimização de hiperparâmetros. Sua proposta é unificar os paradigmas de otimização com base em três princípios fundamentais: design de API orientado pela execução, implementação eficiente, facilidade de configuração e versatilidade de arquitetura (AKIBA et al., 2019) (OPTUNA, 2023). Essa abordagem visa superar as limitações de outras ferramentas, proporcionando uma solução mais integrada e adaptável às diversas demandas, sem comprometer a escalabilidade.

O Optuna usa um algoritmo de busca em árvore baseado em amostras chamado *Tree-structured Parzen Estimator* (TPE) para encontrar os melhores hiperparâmetros. O TPE é um algoritmo de otimização de hiperparâmetros baseado em modelo que usa um modelo de densidade de probabilidade para modelar a distribuição dos hiperparâmetros. O algoritmo divide o espaço de hiperparâmetros em duas partes: uma parte para os hiperparâmetros promissores e outra para os hiperparâmetros não promissores. Em seguida, ele usa o modelo de densidade de probabilidade para encontrar a próxima configuração de hiperparâmetros a ser avaliada. O TPE é altamente eficiente e pode encontrar os melhores hiperparâmetros em menos iterações do que outros algoritmos de otimização de hiperparâmetros.

No Algoritmo 1 é possível ver um exemplo de otimização de hiperparâmetros usando Optuna e TensorFlow. Antes de começar a implementar a otimização com Optuna, é necessário definir uma função objetivo. A função objetivo conterá toda a lógica de um processo regular de definição, treinamento e teste de modelo. Após a avaliação do modelo, ele deve retornar a métrica de avaliação que também é escolhida pelo usuário. A classe Trial será usada para armazenar informações de uma combinação específica de hiperparâmetros usados posteriormente pelo modelo de aprendizado de máquina. Um objeto de estudo pode então ser chamado para otimizar a função objetivo para encontrar a melhor combinação de hiperparâmetros. Em seguida, ele executará testes iterativamente até um teste ou tempo máximo definido pelo usuário. O ensaio com os melhores hiperparâmetros será armazenado em *study.best_trial* (PINHEIRO; BECKER, 2023).

Algoritmo 1: Otimização de Hiperparâmetros com Optuna e TensorFlow

```
import tensorflow as tf
import optuna

# Define a função objetivo a ser maximizada.
def objective(trial):
    # Sugere valores para os hiperparâmetros usando um objeto de teste.
    n_layers = trial.suggest_int('n_layers', 1, 3)
    model = tf.keras.Sequential()
    model.add(tf.keras.layers.Flatten())
    for i in range(n_layers):
        num_hidden = trial.suggest_int(f'n_units_l{i}', 4, 128, log=True)
        model.add(tf.keras.layers.Dense(num_hidden, activation='relu'))
    model.add(tf.keras.layers.Dense(CLASSES))
    # ... Lógica adicional para treinamento e avaliação do modelo ...
    return accuracy

# Cria um objeto de estudo e otimiza a função objetivo.
study = optuna.create_study(direction='maximize')
study.optimize(objective, n_trials=100)
```

2.5 Considerações finais

Este capitulo teve como objetivo fornecer informações sobre o IEEE 802.11, padrão abordado no desenvolvimento deste trabalho. Descreveu as características de RNAs e LSTM, modelos que serão utilizados neste experimento. As últimas seções deste capítulo teve o foco em descrever algumas formas de coletas de dados, indispensável para esse trabalho para a limpeza e pré-processamento que será utilizada como entradas para os modelos e a descrição de um algoritmo de otimização de hiperparâmetros para minimizar o erro quadrático médio do modelo de predição.

Tendo em vista que a proposta do trabalho é coletar os dados de uma rede IEEE 802.11 e fazer o pré-processamento, é necessário o entendimento do funcionamento dos modelos para que seja possível adaptá-los de acordo com os atributos extraídos da limpeza de dados.

3 Trabalhos Relacionados

Neste capítulo são apresentados os trabalhos relacionados a essa pesquisa, concentrando-se na previsão de carga de usuários em rede sem fio.

A predição de carga de usuários se baseia na análise de algumas características da rede, identificando correlações como tendências e períodos sazonais. Por isso, é indispensável a investigação de registros históricos do tráfego de dados passados.

Além disso, é necessário modelos matemáticos e computacionais de previsão baseados em séries temporais para a técnica de predição. Existem técnicas desde as mais simples até as mais complexas, com suas vantagens e desvantagens, que estão relacionadas à precisão da predição, complexidade de implementação, complexidade computacional, tempo de predição, entre outros.

3.1 RNA feedforward multicamada para previsão de carga de usuário em sistemas de conexão sem fio

No trabalho de Abinoja et al. (2015) é proposto e implementado uma RNA *feedforward* multicamada baseada no algoritmo de treinamento Levenberg- Marquardt para estimar a carga do usuário em sistemas de conexão sem fio sem acessar o próprio ponto de acesso. Este algoritmo de otimização procura o mínimo local em uma função e converge mais rapidamente do que um algoritmo genético, tendo maior eficiência para atualizar os pesos da rede neural.

Para estas estimativas, as taxas de dados e a intensidade do sinal foram alimentadas como uma entrada para a rede neural. Após o treinamento, a rede foi testada colocando diferentes entradas nos dados. As saídas da rede neural são comparadas com as saídas reais e em seguida é calculado os erros da rede. Foi utilizado um único modelo de camada oculta com 21 neurônios na sua arquitetura. O melhor desempenho de validação registrado foi de um erro mínimo de 0,111 pela métrica erro quadrático médio (*mean*

squared error - MSE), concluindo que a utilização de rede funcionou bem para o seu propósito. A Figura 3.1 mostra a implementação da rede neural artificial para o sistema do trabalho mencionado.

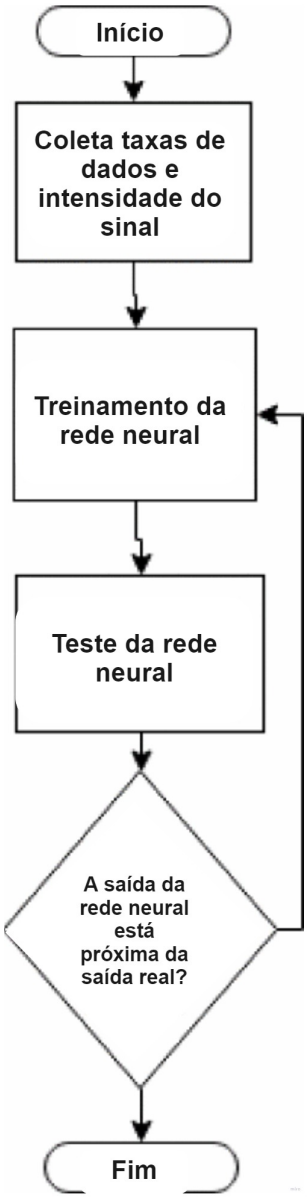

Figura 3.1: Implementação da RNA (ABINOJA et al., 2015)

Ao final do trabalho, é evidenciado como a aplicação de técnicas de redes neurais é vantajosa, o treinamento da rede neural também permite customização e modelagem rápida e fácil em comparação com a modelagem manual, algo que em modelos que possuem muitas variáveis estranhas no processo de treinamento da RN, como camadas ocultas, é impraticável.

3.2 Técnicas e análises de inteligência computacional para auxílio em soluções de dimensionamento de tráfego

Neste trabalho, Frank et al. (2021) apresenta uma aplicação aplicando aprendizado de máquina e uma técnica de otimização para capacitar um ecossistema inteligente. Ele expõe um dos grandes problemas ocasionados pelo aumento da densidade populacional, os engarrafamentos, e apresenta os usos de técnicas e análises de inteligência computacional para estimar o volume de tráfego para apoiar a administração no processo de tomada de decisão.

Para validar, foi realizada uma avaliação utilizando o *Perceptron multicamadas* (MLP) combinado com Otimização por Enxame de Partículas (PSO) com o objetivo de otimizar os parâmetros de uma rede neural artificial com três camadas ocultas. As avaliações foram feitas usando tráfego de dados reais no cenário de fluxo de tráfego livre. Os parâmetros otimizados foram o número de neurônios e a taxa de aprendizagem. O modelo proposto é composto por RNA MLP com três camadas e não utiliza apenas observações de dados recentes, mas também observações históricas, analisando o volume de tráfego do mesmo dia e horário das semanas e anteriores.

Os resultados obtidos após a utilização dessas técnicas de otimização dos parâmetros do modelo, a avaliação se deu pela métrica Média Percentual Absoluta do Erro (MAPE), que toma o valor absoluto do erro percentual entre os valores observados e previstos para cada unidade de tempo e, em seguida, calcula esses valores. O MAPE alcançou um valor mínimo de 3.1% para a função de ativação logística. Continuando com resultados melhores que os obtidos na literatura em que ele faz a comparação e mostrando que a utilização da técnica PSO foi eficaz ao melhorar os resultados do cenário de fluxo livre.

3.3 FLAG: Previsão de carga de usuário flexível, precisa e de longo prazo em sistema Wi-Fi de grande escala usando Deep RNN

Em (CHEN et al., 2021) os autores visam apresentar uma abordagem inovadora para antecipar a carga de usuários em sistemas Wi-Fi de grande escala, almejando uma previsão flexível, precisa e de longo prazo. A carga de usuários, definida como o número de usuários associados a um AP em um instante específico, desempenha um papel crucial na otimização do gerenciamento de recursos, alocação de espectro e qualidade de serviço em sistemas Wi-Fi.

O método proposto, denominado FLAG (*Flexible, Accurate, and Long-Time User Load Prediction*), destaca-se por sua capacidade de personalização em termos de granularidade temporal e horizonte de previsão. Esta flexibilidade permite a previsão da carga de usuários em intervalos de 5 minutos, 15 minutos ou 1 hora, abrangendo horizontes temporais de 30 minutos, 1 hora ou 24 horas.

O FLAG é composto por três elementos fundamentais: aquisição de dados, extração de características e design do modelo. Na fase de aquisição de dados, mais de 25 milhões de registros de associação provenientes de mais de 55 mil usuários foram processados pelos autores, visando extrair a essência da carga de usuários em nível de AP. Quanto à extração de características, uma análise abrangente foi realizada para identificar elementos cruciais que foram classificados em três categorias: características individuais, características espaciais e características temporais.

As características individuais descrevem propriedades específicas de cada AP, como o número médio de usuários, variância da carga de usuários e taxa de ocupação. Já as características espaciais capturam a correlação entre APs vizinhos, incluindo distância, similaridade e interferência. Por fim, as características temporais refletem a variação da carga de usuários ao longo do tempo, incorporando elementos como tendência, sazonalidade e periodicidade.

No que tange ao design do modelo, os autores propuseram uma rede neural recorrente profunda, composta por duas RNNs distintas: a RNN codificadora e a RNN

decodificadora. A RNN codificadora utiliza vetores de características sequenciais de cada AP como entrada, aprendendo assim uma representação semântica da carga de usuários. Essa representação é então utilizada pela RNN decodificadora para gerar previsões sequenciais da carga de usuários para cada AP. Destaca-se que a representação semântica é injetada em cada etapa de previsão, reduzindo eficazmente os erros acumulados e possibilitando previsões de longo prazo.

Os autores implementaram uma instância do FLAG em um sistema Wi-Fi operacional com mais de 7000 APs e conduziram experimentos com dados reais para avaliar sua eficácia. Os resultados demonstraram que o FLAG superou diversos métodos de previsão existentes em termos de precisão, flexibilidade e escalabilidade. Ademais, foram evidenciadas aplicações potenciais do FLAG, tais como o balanceamento de carga, seleção de AP e detecção de anomalias.

3.4 Controle AP inteligente em grande escala com notável economia de energia no sistema Wi-Fi do campus

No trabalho de (FANG et al., 2018) os autores propõem um esquema de controle inteligente de pontos de acesso (APs) em larga escala, chamado ACE (Controle de AP com economia de energia), para controlar dinamicamente os APs em larga escala (ligados ou desligados para economia de energia, sem perda de cobertura Wi-Fi). O esquema ACE é inspirado em dados de status de AP em larga escala coletados no sistema Wi-Fi do campus, que contém mais de 8.000 APs e atende cerca de 40.000 usuários finais ativos em uma área de 3,0925 km^2. Depois de conduzir estudos empíricos sobre as cargas de AP, os autores descobriram que o fenômeno ocioso prevalece em todo o rastreamento. Uma grande parte dos APs está funcionando sem nenhuma associação de usuário, o que inevitavelmente levará a um consumo desnecessário de energia. Para resolver esse problema, o esquema ACE usa o algoritmo de floresta aleatória para prever a carga de cada AP e desliga aqueles cujas durações ociosas duram mais do que o comprimento da janela deslizante pré-definida.

O ACE integra dois componentes principais: modelagem e previsão de carga de AP; algoritmo de controle AP em larga escala. Foram conduzidas extensas simulações baseadas em rastreamento para avaliar o desempenho do esquema ACE e os resultados demonstraram sua eficiência.

Por fim, os autores conduzem simulações extensivas baseadas em traços para demonstrar a eficiência do esquema ACE; especificamente, mais de 70% da energia pode ser economizada com mais de 92% da cobertura Wi-Fi do usuário garantida em média 1.

3.5 Modelos de aprendizado de máquina que exploram conjuntamente as correlações espaço-temporais

Em Qiu et al. (2018), os autores avaliaram a predição de carga de tráfego de dados em redes sem fio utilizando uma relação espaço temporal e rede neural recorrente, como *Gated Recurrent Unit* (GRU), que visa resolver o problema da dissipação do gradiente que é comum em uma rede neural recorrente padrão, e *Long Short Term Memory* (LSTM), uma arquitetura de rede neural recorrente que "lembra" valores em intervalos arbitrários. Nele o autor propõe alguns modelos de aprendizado de máquina baseados em múltiplas Redes Neurais Recorrentes, que exploram conjuntamente as correlações espaço temporais com objetivo de melhorar o desempenho da previsão de tráfego.

Além disso, ele apresenta uma abordagem de aprendizado multitarefa que é uma maneira promissora de melhor o aprendizado e prever o desempenho, considerando conjuntamente várias entradas enquanto os diferentes recursos entre as tarefas podem ser utilizados de forma eficaz, isso leva à resolução de tarefas ou problemas mais complexos no menor tempo possível usando o aprendizado de máquina.

Para o ajuste dos hiperparâmetros, eles usam o algoritmo de otimização bayesiana para os métodos de comparação e apresentar os resultados otimizados, e observam que os modelos baseados em RNN superam todas as abordagens de comparação na maioria dos casos.

Os resultados, quando a RNN não possui neurônios suficientes, a capacidade de representação da informação é limitada, especialmente para o caso da arquitetura n-para-

n, onde informações sobrecarregadas causam *underfitting*. Ao aumentar o número de neurônios, mais recursos são extraídos. No entanto, a melhoria é interrompida depois que o tamanho de 150 neurônios é atingido na máquina de aprendizado.

Com base em dados reais, os autores forneceram avaliação detalhada em diferentes modelos de aprendizado e demonstraram que a correlação espacial entre as estações base pode fornecer informações valiosas para melhorar a precisão da previsão.

3.6 Considerações finais

No presente capítulo, apresentou-se uma visão abrangente dos trabalhos relacionados que embasam e contextualizam o desenvolvimento do método proposto. Estas pesquisas oferecem contribuições substanciais para a otimização de soluções relacionadas ao dimensionamento de tráfego e à previsão de carga, refletindo avanços significativos na eficiência operacional de redes sem fio.

Esses estudos desempenham papéis cruciais, fornecendo alicerces essenciais para a concepção do trabalho em questão. As contribuições destacadas enriqueceram significativamente a compreensão do panorama de previsão de carga de usuários em sistemas Wi-Fi de grande escala, exercendo influência direta no refinamento e aprimoramento do método proposto.

4 Proposta

Neste capítulo é realizado uma descrição da metodologia utilizada no modelo de Redes Neurais com o objetivo de fazer a predição de carga de usuários.

Conforme já destacado, para o desenvolvimento do trabalho, uma das etapas principais e necessárias foi a avaliação correta do estado da arte, juntamente com uma pesquisa bibliográfica muito seletiva com enfoque em inteligência artificial e redes de larga escala. Durante esse processo, o principal objetivo era encontrar os modelos de aprendizagem profunda mais relevantes e populares que são usados na previsão de dados.

Uma escolha particularmente apropriada para lidar com séries temporais, como a quantidade de autenticações de usuários por hora em uma rede Wi-Fi, é o uso de Redes Neurais LSTM. As LSTMs são capazes de capturar dependências de longo prazo nos dados sequenciais, o que se revela fundamental ao lidar com previsões em séries temporais (HOCHREITER; SCHMIDHUBER, 1997).

4.1 Base de dados

Nesta seção, é descrito toda a compreensão da base de dados que serviu como fundação para o treinamento do modelo de Redes Neurais. A escolha cuidadosa da fonte de dados e os procedimentos de coleta e processamento são elementos cruciais para a eficácia do modelo na predição de carga de usuários em redes Wi-Fi.

4.1.1 Coleta de dados

Os dados utilizados para o treinamento do modelo foram coletados de uma base de dados reais. A base de dados investigada tem como origem os *logs* gerados pela controladora da rede Wi-Fi eduroam no Brasil (EDUROAM, 2022). Essa controladora está localizada na infraestrutura da Rede Nacional de Ensino e Pesquisa (RNP) (Rede Nacional de Ensino e Pesquisa, 2022).

4.1.2 Processamento dos dados

No arquivo de *logs* gerado que contém informações sobre todas as autenticações do servidor, existem diversas informações que não são importantes para esse trabalho proposto, portanto foi preciso fazer um tratamento dos dados, para que remanesça os dados significantes, os quais possuem maior quantidade de atributos de relevância para a pesquisa.

Para restringir o escopo do experimento, conduziu-se uma pesquisa na base de dados a fim de extrair informações relevantes sobre as autenticações aceitas por usuários distintos a cada hora, abrangendo o período de 2019 a setembro de 2023. A instituição escolhida para essa análise foi a Universidade de São Paulo (USP), selecionada devido ao considerável volume diário de autenticações. A Tabela 4.1 apresenta um fragmento representativo da base de dados, compilando as informações extraídas dos registros de *logs*. Esse procedimento tem como objetivo primário delimitar o escopo da pesquisa para a USP, proporcionando uma base inicial que, posteriormente, poderá ser expandida para um contexto nacional.

Tabela 4.1: Exemplo da base de dados utilizada.

id	datetime	count
0	2019-01-01 00:00:00	20
1	2019-01-01 01:00:00	38
2	2019-01-01 02:00:00	41
3	2019-01-01 03:00:00	50
4	2019-01-01 04:00:00	61

No contexto do experimento, foi identificado que a base de dados apresentava lacunas significativas nos registros de séries temporais nos meses de fevereiro de 2019 e junho de 2020, como mostra a Figura 4.1. Essas lacunas poderiam comprometer a integridade e confiabilidade da análise, uma vez que a ausência de dados pode distorcer as tendências e padrões observados.

Diante desse desafio, optou-se por adotar uma estratégia de preenchimento das lacunas, visando minimizar o impacto dessas ausências nos resultados finais. No entanto, ao analisar mais profundamente, constatou-se que os anos de 2020 e 2021 foram fortemente afetados pela pandemia global, resultando em padrões atípicos de comportamento e autenticações.

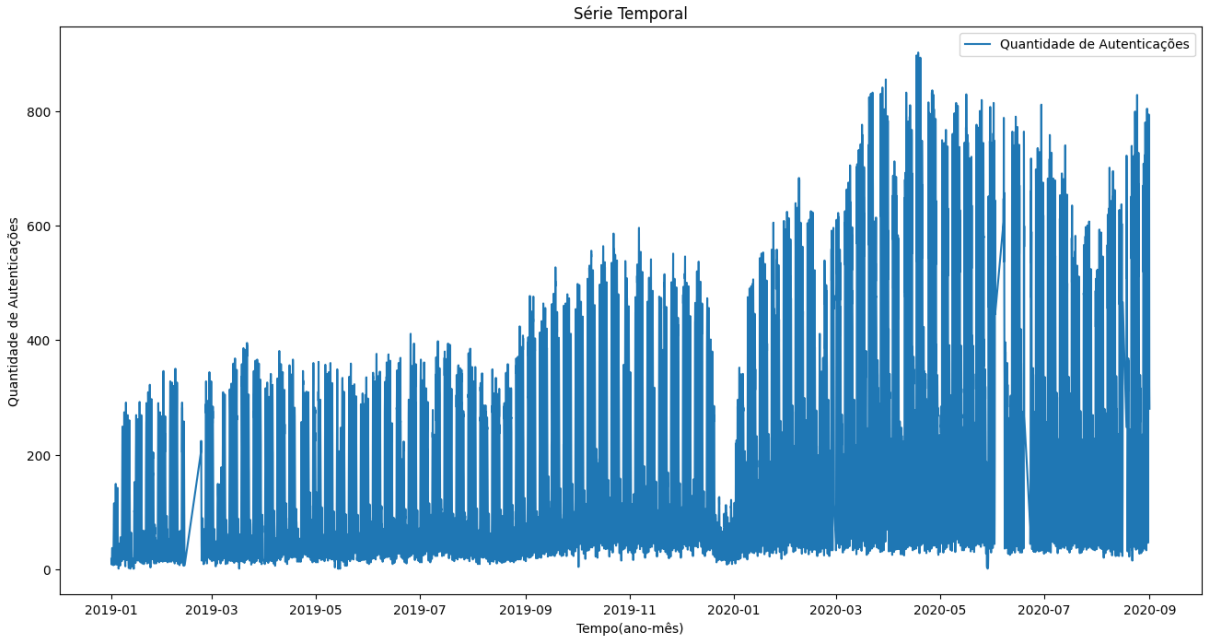

Figura 4.1: Série temporal da base utilizada
Elaborado pela autora. (2023)

Considerando a necessidade de manter a consistência e representatividade dos dados, decidiu-se descartar os anos de 2020 e 2021 da análise. Dessa forma, o ano de 2020 foi excluído da investigação, e os dados correspondentes a 2023 foram considerados em seu lugar. Essa abordagem teve como objetivo assegurar que as conclusões obtidas refletissem de maneira mais precisa a realidade das autenticações, evitando distorções causadas pelas circunstâncias extraordinárias dos anos afetados pela pandemia.

Para a implementação da estratégia de preenchimento de lacunas, primeiramente, desenvolveu-se um índice abrangente que engloba o período compreendido entre 2019 e agosto de 2020, assegurando a representação de todas as horas no referido intervalo temporal. Subsequentemente, adotou-se a abordagem da média móvel ponderada para preencher os valores ausentes, visto que ela emprega considerações acerca das tendências temporais ao longo de um determinado intervalo de tempo, assegurando a representatividade e realismo dos novos valores gerados. Para cada valor carente, calculou-se uma média, levando em consideração os valores nos mesmos horários em dias anteriores, dentro de uma janela de sete dias para a média móvel.

Após o preenchimento das lacunas, procedeu-se ao arredondamento dos valores para números inteiros, simplificando, assim, o conjunto de dados. Por fim, visando manter

a consistência da coluna de identificação, esta foi reorganizada de modo a iniciar em 0 e aumentar sequencialmente.

O resultado alcançado consiste em um conjunto de dados mais completo e preparado para análise, proporcionando a manipulação de informações mais consistentes e organizadas, como mostra a Figura 4.2.

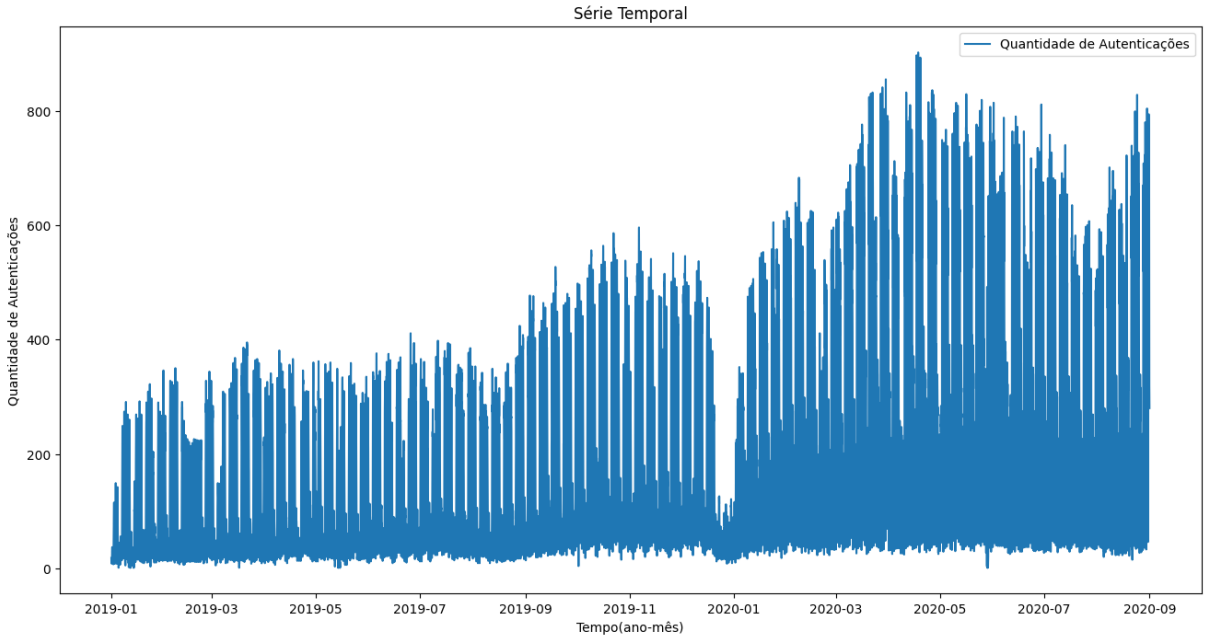

Figura 4.2: Série temporal da base utilizada com lacunas preenchidas
Elaborado pela autora. (2023)

4.2 Estrutura do modelo

No processo de implementação das redes neurais para o desenvolvimento do modelo, foi usado a linguagem de programação Python 3.9[1]. Já as bibliotecas utilizadas, foram escolhidas a Keras[2] que possui código aberto e de forma simples, e a TensorFlow [3], uma poderosa biblioteca de código aberto para aprendizado de máquina e redes neurais. Reconhecida por sua eficiência e flexibilidade, a TensorFlow contribui para a construção de modelos complexos e oferece recursos avançados para treinamento e avaliação.

Para aprimorar ainda mais o processo de pré-processamento e otimizar a ava-

[1]https://www.python.org/
[2]https://keras.io/
[3]https://www.tensorflow.org/

liação do modelo, outras bibliotecas foram incorporadas. Nesse contexto, a scikit-learn[4] desempenha um papel adicional, proporcionando ferramentas essenciais para tarefas como escalonamento de dados, seleção de características e métricas de avaliação.

Dessa forma, a seleção cuidadosa dessas ferramentas e bibliotecas não apenas facilita o desenvolvimento eficiente do modelo de redes neurais, mas também estabelece uma base sólida para análises e melhorias futuras.

Uma etapa crucial no processo de preparação dos dados foi a normalização dos valores. Essa normalização foi realizada por meio do Min-Max Scaler, transformando os dados para um intervalo padronizado entre 0 e 1. Essa normalização é fundamental para o bom desempenho de modelos de redes neurais, garantindo que todas as variáveis estejam na mesma escala.

Com o intuito de preparar os dados para o treinamento do modelo, uma função foi desenvolvida para criar sequências de dados. Cada sequência representa um histórico de valores anteriores, sendo que o valor subsequente é designado como o alvo da previsão. Neste contexto, optou-se por sequências com 24 (vinte e quatro) horas, ou seja, um dia, visando capturar padrões diários de utilização.

Posteriormente, o conjunto de dados foi dividido em conjuntos de treinamento e teste. 80% das sequências foram alocadas para o treinamento do modelo, enquanto os 20% restantes foram reservados para avaliar o desempenho do modelo em dados não vistos durante o treinamento.

A arquitetura do modelo LSTM foi então construída utilizando a biblioteca TensorFlow em conjunto com Keras. O modelo compreende uma camada LSTM, que aceita uma sequência de 37 passos temporais, representando o histórico da série temporal, concebida para capturar padrões temporais complexos presentes nos dados. Entre as camadas LSTM, é aplicada uma técnica de regularização chamada dropout, com uma taxa de aproximadamente 13.8%. Adicionalmente, uma camada densa final com um único neurônio foi incorporada para gerar as previsões.

[4]https://scikit-learn.org/stable/

4.3 Métricas

Diversas métricas de medição de erro são utilizadas para avaliar a qualidade das predições. Para a avaliação e análise do modelo foi utilizado quatro métricas de erros que são comuns na comparação de redes neurais.

Os erros são: Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE), Erro Médio Absoluto (MAE) e Erro Percentual Absoluto Médio (MAPE).

MSE: é uma métrica que calcula a média dos quadrados dos desvios entre as predições do modelo e os valores reais. Sua fórmula matemática é expressa por:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.1}$$

RMSE: é uma extensão do MSE, calculando a raiz quadrada da média dos quadrados dos desvios. Essa métrica proporciona uma interpretação mais intuitiva, representando a média dos desvios numa escala similar à dos dados originais:

$$RMSE = \sqrt{MSE} \tag{4.2}$$

MAE: diferentemente do MSE, este mensura a média dos valores absolutos dos desvios entre predições e valores reais. O MAE é menos sensível a *outliers*, ou seja, valores atípicos que se destacam significativamente em relação aos demais no conjunto de dados. *Outliers* podem distorcer as métricas de avaliação, como o MSE, que dá maior ênfase a grandes desvios. No entanto, o MAE, ao calcular a média dos valores absolutos dos desvios, proporciona uma avaliação mais robusta, minimizando o impacto de observações extremas na precisão geral do modelo. Sua fórmula é dada por:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.3}$$

MAPE: expressa os desvios como uma porcentagem da magnitude real dos valores. Ele é particularmente útil para interpretar a precisão do modelo em termos percentuais, proporcionando uma compreensão mais intuitiva da performance.Sua formulação é

dada por:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100 \tag{4.4}$$

O treinamento de cada modelo foi repetido cinco vezes, de forma que as métricas finais representam a média dos cinco modelos produzidos.

5 Avaliação e Resultados

Este capítulo se dedica à avaliação do modelo de Redes Neurais LSTM proposto, explorando as métricas empregadas para mensurar sua eficácia na predição da carga de usuários em redes Wi-Fi. Ao longo do processo, serão apresentados os resultados obtidos a partir da análise das bases de dados e do modelo de predição utilizado. Além disso, será examinado o desempenho do modelo, seguido por uma análise concisa das descobertas.

5.1 Cenário de Testes

A Tabela 5.1 mostra os cenários que foram realizados os testes para LSTM. Durante o treinamento, foram utilizados lotes de 40 exemplos, uma camada com 86 neurônios e o modelo foi treinado ao longo de 82 épocas. Esses valores específicos foram escolhidos após uma otimização sistemática realizada pelo otimizador Optuna, que busca automaticamente os melhores hiperparâmetros para maximizar o desempenho do modelo na tarefa de previsão de séries temporais. O Optuna realiza uma busca inteligente no espaço de hiperparâmetros, ajustando-os iterativamente para minimizar o MSE, que foi a função objetivo escolhida. Essa abordagem de otimização resultou nos parâmetros aqui mencionados, que foram considerados como os mais eficazes para o conjunto de dados em questão.

Tabela 5.1: Cenário de teste para LSTM

	Cenário	**Otimizador**	**Função Ativação**	**Número de neurônios**
1	1	Adam	Relu	100, 200, 500
2	2		Sigmoide	

5.2 Resultados dos Experimentos

Para cada cenário disponível foram executados 5 treinamentos de cada modelo, de forma que os valores das métricas de erro analisadas são as médias das execuções.

- **Cenário 1:** cenário da arquitetura LSTM, com a utilização do algoritmo de otimização Adam e função de ativação Relu.

Tabela 5.2: Cenário de teste para LSTM - Relu

Erro Médio de Validação				
Número de Neurônios	**MSE**	**RMSE**	**MAE**	**MAPE (%)**
100	1875.3	43.2	27.8	35.3
200	1966.5	43.8	27.8	42.0
500	2091.9	41.7	29.2	42.7

Na Tabela 5.2 estão dispostos os erros médios RMSE, MSE e MAPE das cinco execuções, sendo todos do conjunto de validação. Para o número de neurônios igual a 100, o modelo obteve um MSE de 1875.3, um RMSE de 43.2, um MAE de 27.8 e um MAPE de 35.3. Aumentando o número de neurônios para 200, os resultados não mostraram uma melhoria significativa, com um MSE de 1966.5, RMSE de 43.8, MAE de 27.8 e MAPE de 42.0. Entretanto, ao aumentar ainda mais o número de neurônios para 500, houve um aumento no MSE para 2091.9, enquanto o RMSE diminuiu para 41.7. O MAE aumentou ligeiramente para 29.2, e o MAPE atingiu 42.7.

Dessa forma, pode-se notar que o aumento no número de neurônios não resultou em melhorias consistentes nas métricas de erro. Isso sugere que, para o cenário em questão, um número menor de neurônios pode ser mais eficaz, evitando possíveis problemas de *overfitting*.

- **Cenário 2:** cenário da arquitetura LSTM, com a utilização do algoritmo de otimização Adam e função de ativação Sigmoide.

Tabela 5.3: Cenário de teste para LSTM - Sigmoid

Erro Médio de Validação				
Número de Neurônios	**MSE**	**RMSE**	**MAE**	**MAPE (%)**
100	2560.2	48.5	34.7	38.6
200	2493.0	47.8	34.6	36.7
500	2180.3	44.6	31.4	36.0

Na análise dos resultados que estão na Tabela 5.3, os erros médios foram calculados para as cinco execuções no conjunto de validação. Considerando o número de

neurônios igual a 100, o modelo apresentou um MSE de 2560.2, indicando a média dos quadrados dos erros entre as previsões e os valores reais. O RMSE correspondente foi de 48.5, sugerindo uma raiz quadrada do MSE e uma medida da dispersão dos erros.

Para a mesma configuração com 100 neurônios, o MAE foi de 34.7, representando a média dos valores absolutos dos erros. O MAPE, que calcula a média das porcentagens dos erros em relação aos valores reais, foi de 38.6 para esse cenário específico.

Ao aumentar o número de neurônios para 200, observou-se uma redução nos erros. O MSE diminuiu para 2493.0, o RMSE para 47.8, o MAE para 34.6 e o MAPE para 36.7. Essa tendência de melhoria nos indicadores de erro continuou com 500 neurônios, resultando em um MSE de 2180.3, RMSE de 44.6, MAE de 31.4 e MAPE de 36.0.

Em resumo, os resultados revelam consistentemente que o aumento no número de neurônios na camada LSTM está associado a uma diminuição nos erros médios, indicando uma melhoria na capacidade do modelo em prever a quantidade de usuários por hora em uma rede, quando utilizando a função de ativação Sigmoide. Ao utilizar a função de Ativação Relu, a diminuição no número de neurônios na camada LSTM já era mais benéfica, visto que com aumento não foram registrados resultados satisfatórios.

5.2.1 Comparação

A comparação entre os dois cenários de teste mais promissores, utilizando as funções de ativação Relu e Sigmoid em uma arquitetura LSTM, revela nuances interessantes sobre o desempenho dos modelos na previsão da quantidade de usuários por hora em uma rede.

No cenário 1, onde a função de ativação Relu foi empregada, observamos que, para um número de neurônios igual a 100, o modelo apresentou um MSE de 1875.3, RMSE de 43.2, MAE de 27.8 e MAPE de 35.3. Aumentar o número de neurônios para 500 resultou em um aumento no MSE para 2091.9, RMSE diminuído para 41.7, MAE levemente aumentado para 29.2 e MAPE atingindo 42.7. Esses resultados sugerem que, no caso da função Relu, um número menor de neurônios (100) foi mais eficaz, evitando

possíveis problemas de *overfitting*.

Por outro lado, no cenário 2, com a função de ativação Sigmoid, observamos uma tendência diferente. Para 100 neurônios, o modelo apresentou um MSE de 2560.2, RMSE de 48.5, MAE de 34.7 e MAPE de 38.6. Aumentar o número de neurônios para 500 resultou em uma melhoria consistente em todas as métricas, com MSE de 2180.3, RMSE de 44.6, MAE de 31.4 e MAPE de 36.0.

Comparando os dois cenários, nota-se que, enquanto na função Relu um número menor de neurônios foi mais eficaz, na função Sigmoid houve uma melhoria significativa ao aumentar o número de neurônios. Isso sugere que a escolha da função de ativação pode impactar a influência do número de neurônios na performance do modelo. Cada função de ativação possui propriedades distintas, e a escolha entre elas deve considerar a natureza específica do problema e dos dados.

Em suma, a comparação entre Relu e Sigmoid destaca a importância de ajustes finos na arquitetura da rede neural, considerando não apenas o número de neurônios, mas também a função de ativação, para otimizar o desempenho do modelo na tarefa de previsão de usuários em uma rede.

Na Figura 5.1, as previsões da LSTM no Cenário 1, que obteve métricas superiores, acompanham de perto o comportamento da série original. Já na Figura 5.2, os resultados da LSTM no Cenário 2, com a função de ativação Sigmoid, também demonstram uma consistência notável com a série original. Ambas as visualizações destacam a eficácia das redes LSTM na captura de padrões temporais complexos e na realização de previsões alinhadas com os dados reais, validando as escolhas arquiteturais específicas de cada cenário.

Nas Tabelas 5.4 e 5.5 são apresentados alguns valores dos resultados preditos da melhor execução considerando a execução com menor valor de MAPE, em comparação com os valores reais.

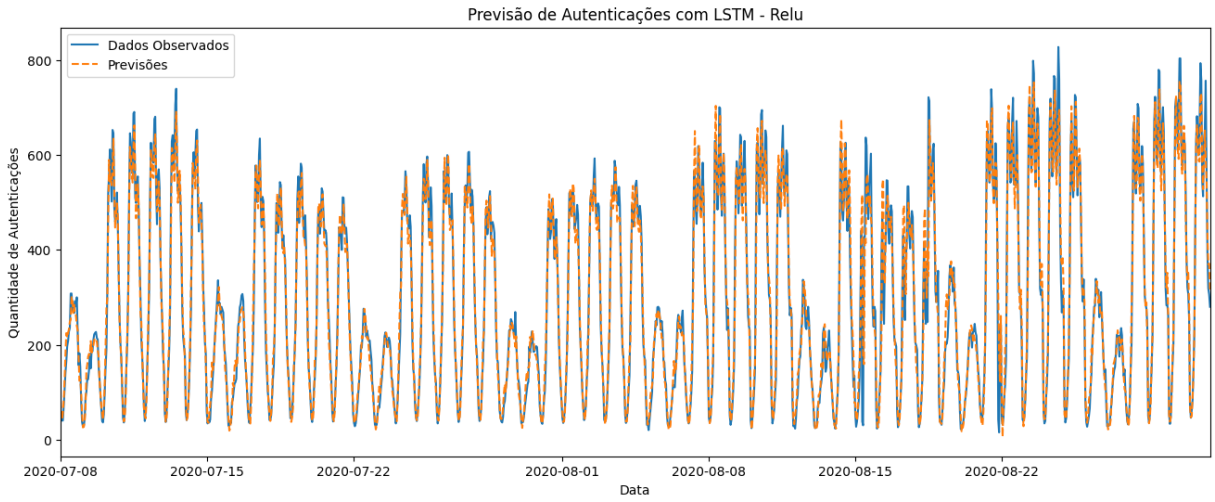

Figura 5.1: Valores preditos pela LSTM com função de ativação Relu

Elaborado pela autora. (2023)

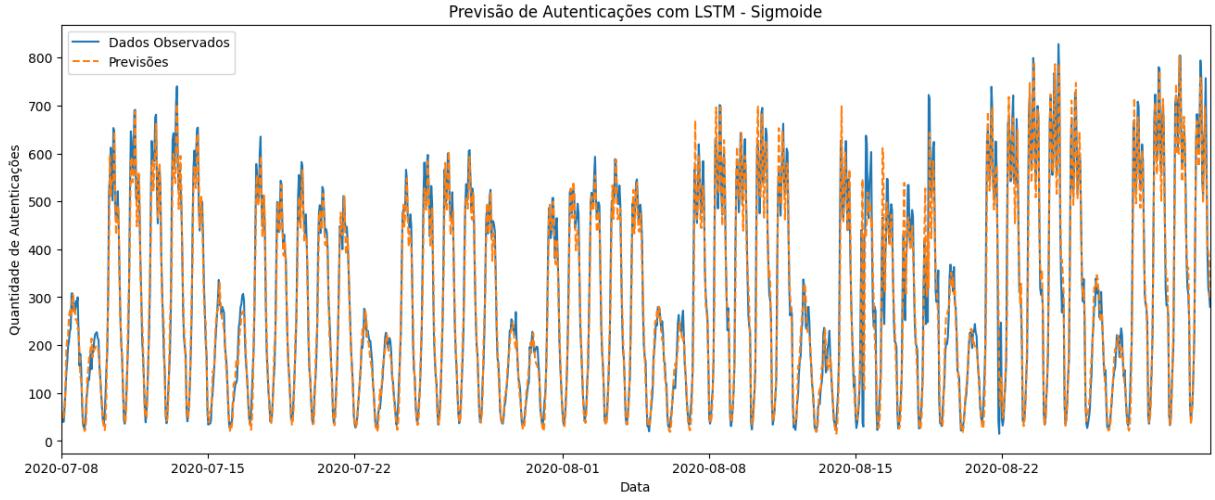

Figura 5.2: Valores preditos pela LSTM com função de ativação Sigmoid

Elaborado pela autora. (2023)

Tabela 5.4: Comparação dos resultados preditos com os valores reais do Cenário 1

ID	**Data**	**Quantidade Previsto**	**Quantidade Real**
0	2020-05-02 12:00:00	715	707
1	2020-05-02 13:00:00	688	666
2	2020-05-02 14:00:00	524	557
3	2020-05-02 15:00:00	532	508
4	2020-05-02 16:00:00	584	555
5	2020-05-02 17:00:00	577	652
...			
1453	2020-08-31 19:00:00	521	568
1454	2020-08-31 20:00:00	386	394
1455	2020-08-31 21:00:00	343	319
1456	2020-08-31 22:00:00	357	306
1457	2020-08-31 23:00:00	293	280

Tabela 5.5: Comparação dos resultados preditos com os valores reais do Cenário 2

ID	Data	Quantidade Previsto	Quantidade Real
0	2020-05-02 12:00:00	689	707
1	2020-05-02 13:00:00	665	666
2	2020-05-02 14:00:00	546	557
3	2020-05-02 15:00:00	510	508
4	2020-05-02 16:00:00	553	555
5	2020-05-02 17:00:00	617	652
...			
1453	2020-08-31 19:00:00	555	568
1454	2020-08-31 20:00:00	392	394
1455	2020-08-31 21:00:00	387	319
1456	2020-08-31 22:00:00	382	306
1457	2020-08-31 23:00:00	315	280

Portanto, com base nos resultados apresentados, a configuração com 100 neurônios e a função de ativação Relu parece ser a mais adequada para o cenário abordado. Essa escolha não apenas demonstrou um bom desempenho nas métricas avaliadas, mas também evitou problemas de overfitting. Já em relação as métricas, se a ênfase for na precisão percentual, o MAPE pode ser uma métrica importante a ser considerada, pois fornece uma medida relativa de precisão que considera a escala dos valores preditos e reais. No entanto, é aconselhável avaliar múltiplas métricas e considerar a interação entre elas para obter uma compreensão mais abrangente do desempenho do modelo.

5.3 Gerência à redes

Nesta seção apresentamos a análise sobre um cenário de aplicação prática da predição de carga para orientar a contratação de um *link*/enlace de internet adequado, e sua correlação na otimização de diversos aspectos na gestão da rede, tais como: economia

financeira, melhoria na qualidade de serviço, prevenção de interrupções, entre outros.

5.3.1 Prevenção de colapso na rede

Ao analisar a predição de carga de autenticações na USP através do modelo LSTM, a necessidade de aplicar esses resultados de forma prática torna-se evidente, especialmente ao considerar o planejamento para os próximos períodos.

Nesse trabalho foi considerado quatro tipos de consumo distintos na rede - videochamadas, redes sociais, para streaming de música e consumo de vídeos online - como representado na Figura 5.3. Para cada uma dessas atividades, é determinado um valor mínimo de velocidade de conexão para garantir uma experiência de usuário satisfatória.

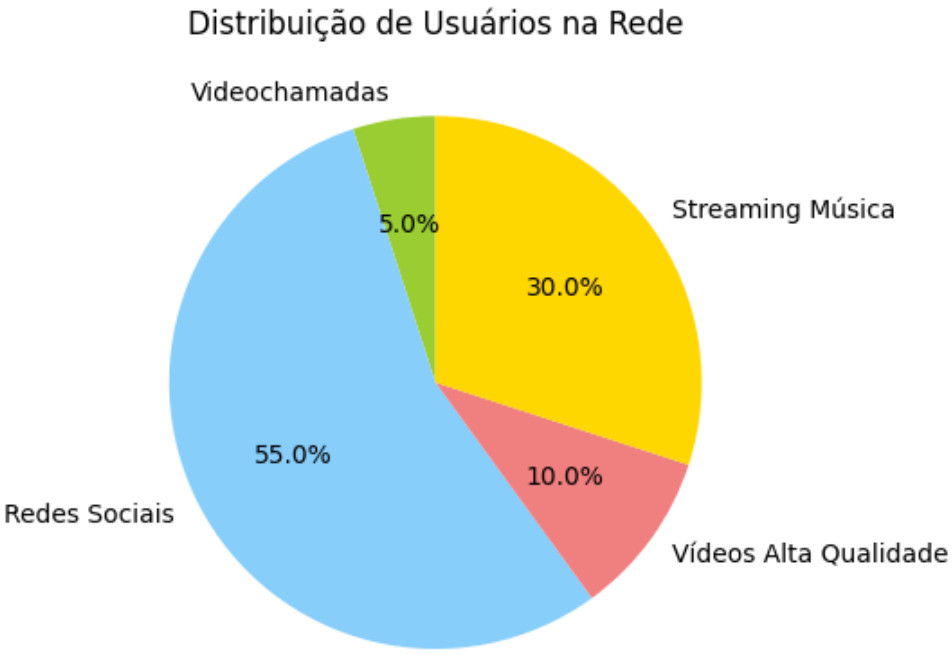

Figura 5.3: Distribuição de usuários na rede
Elaborado pela autora. (2023)

Conforme estabelecido no Artigo 2º, inciso II, alínea a, da Portaria nº 33, de 7 de agosto de 2023, que define os critérios da Política de Inovação Educação Conectada (BáSICA, 2023), fica determinado que o serviço de acesso à internet nas escolas de educação básica deve apresentar uma velocidade mínima de 1 Mbps por estudante no maior turno.

Segundo a Nasa Tecnologia (NASA...,), referência nacional em cabeamento

estruturado de redes para empresas, os planos de internet de fibra ou links dedicados oferecem velocidades simétricas, o que significa que suas velocidades de upload e download serão as mesmas. Assim, se tiver uma conexão de fibra de 100 Mbps, sera possível fazer upload de arquivos também a 100 Mbps.

A Tabela 5.6 mostra a quantidade de Mbps necessária por dispositivo para atividades comuns na Internet.

Tabela 5.6: Mbps necessária por dispositivo para atividades comuns na Internet

Atividades	**Mínimo (Mbps)**	**Recomendada (Mbps)**
E-mail	1	2
Navegação na web (Sites, Redes Sociais, etc)	5	15
Transmissão de Vídeo HD	10	25
Transmissão de Vídeo 4K	25	100
Streaming de Música	1	5
Videochamadas Individuais	5	25
Telefones VoIP	1	1

A identificação do dia com o maior volume de autenticações foi realizada, revelando que em 18 de março de 2020, às sete horas da manhã, ocorreram 902 autenticações. Esse dado pode ser visto na Figura 5.4 Esses dados permitem avaliar se o link dedicado da USP é suficiente para acomodar essa quantidade de usuários.

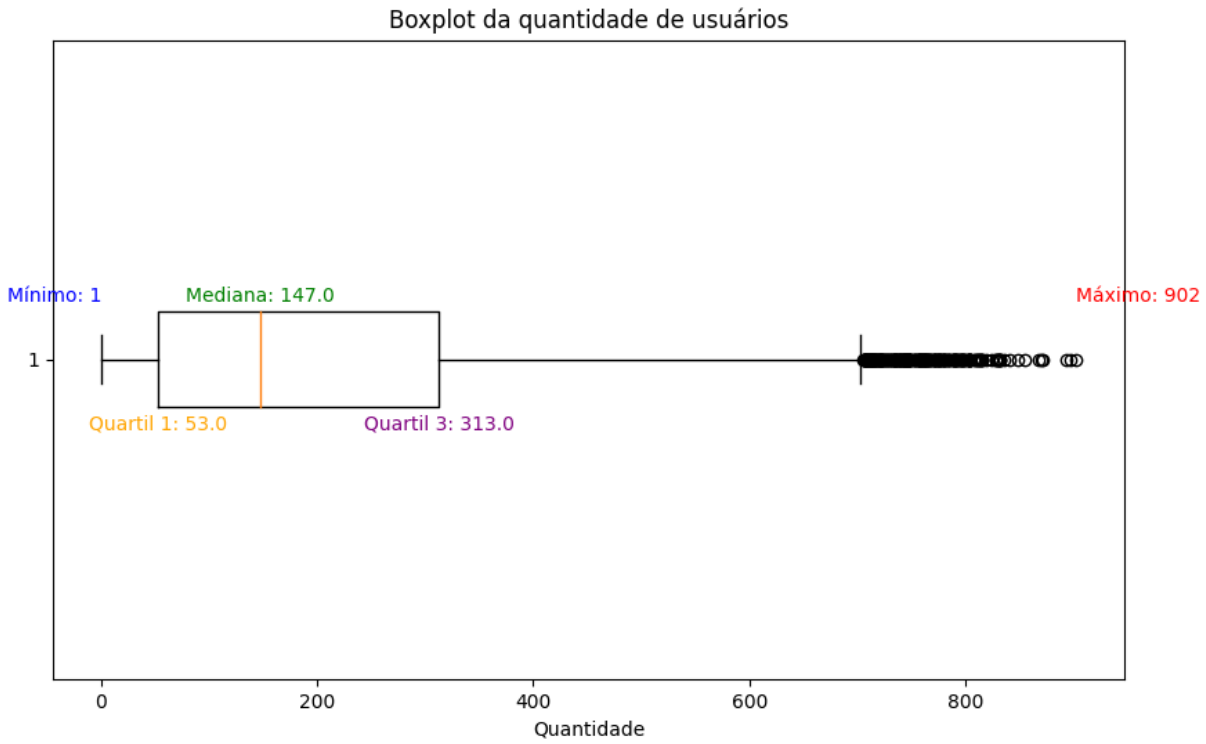

Figura 5.4: Limites superior e inferior da quantidade de usuários por hora

Elaborado pela autora. (2023)

Assim, considerando que a USP conta com até 902 estudantes no maior turno, a contratação de um plano de 900 Mbps atende ao parâmetro de 1 Mbps por estudante no maior turno, conforme preconizado. Resta a indagação se essa capacidade é adequada para suportar os estudantes envolvidos em diversas atividades online.

Para sanar essa indagação, alguns cálculos foram feitos utilizando como parâmetro a velocidade de internet mínima por cada atividade:

1. Redes Sociais:

$$\text{Gasto\_RS} = (902 \times 0,55) \times 5 \text{ Mbps} = 2479,5 \text{ Mbps}$$

2. Streaming de música:

$$\text{Gasto\_SM} = (902 \times 0,3) \times 1 \text{ Mbps} = 270,6 \text{ Mbps}$$

3. Vídeos online de alta qualidade:

$$\text{Gasto\_VO} = (902 \times 0,1) \times 10 \text{ Mbps} = 901 \text{ Mbps}$$

4. Videochamadas:

$$\text{Gasto\_V} = (902 \times 0,05) \times 5 \text{ Mbps} = 225,5 \text{ Mbps}$$

O gasto total de largura de banda será:

$$\text{Gasto\_Total} = \text{Gasto\_RS} + \text{Gasto\_SM} + \text{Gasto\_VO} + \text{Gasto\_VO}$$

$$\text{Gasto\_Total} = 2479,5 + 270,6 + 901 + 225,5 = 3.876,6 \text{ Mbps}$$

A análise dos cálculos realizados oferece uma visão abrangente do consumo potencial de largura de banda para as diversas atividades online dos estudantes. Ao considerar o uso de redes sociais, streaming de música, vídeos online de alta qualidade e videochamadas, chegamos a um gasto total de largura de banda de 3.876,6 Mbps.

Esse resultado revela que a alocação de uma conexão de 900 Mbps, embora atenda ao requisito mínimo de 1 Mbps por estudante no maior turno, pode não ser suficiente para garantir uma experiência de usuário satisfatória. O consumo projetado supera significativamente a capacidade disponível, indicando a necessidade de considerar planos de internet com maior largura de banda para acomodar as demandas diversificadas dos estudantes em suas atividades online.

Em um cenário prático, a instituição pode utilizar as previsões para otimizar a contratação de links/enlaces de internet. Conhecendo com antecedência os períodos de maior demanda, pode-se adotar contratos elásticos, ajustando a capacidade da conexão de acordo com a necessidade real. Isso representa uma mudança significativa em relação aos contratos estáticos, possibilitando uma gestão mais eficiente e econômica dos recursos de rede.

Dessa forma, a capacidade de prever a carga de autenticações não apenas con-

tribui para a estabilidade e confiabilidade da rede, mas também abre espaço para uma abordagem mais dinâmica e eficiente na contratação de recursos, alinhando-se melhor às demandas reais da instituição. Essa adaptação proativa é fundamental para a prevenção de possíveis gargalos e colapsos, promovendo uma gestão mais eficaz e econômica dos recursos de conectividade.

5.3.2 Economia Financeira

Os contratos de provedores de internet geralmente são baseados no mínimo garantido, que é a quantidade mínima de largura de banda que será fornecida independentemente da demanda. Essa é uma prática comum em contratos de serviços de internet para garantir uma certa qualidade de serviço, especialmente em situações em que há flutuações na demanda ao longo do tempo.

A observação atenta da utilização da rede revela que o pico de 902 usuários não é uma ocorrência frequente, conforme visto na Figura 5.4. Durante a maior parte do dia, a quantidade de usuários conectados é substancialmente menor, com 700 usuários sendo um valor mais comum. Diante dessa realidade, a estratégia proposta consiste em calcular a média dos usuários conectados durante os períodos em que a demanda é inferior ao pico.

Ao calcular a média dos usuários conectados, especialmente durante os momentos de menor demanda, podemos determinar um valor mais realista para a largura de banda mínima necessária. Essa abordagem reconhece que a infraestrutura da rede não precisa ser dimensionada para atender ao pico raro, mas sim para manter uma qualidade de serviço satisfatória durante a demanda média.

A média dos usuários na base analisada é de aproximadamente 200.02, já a largura de banda total para os 200 usuários é mostrado a seguir.

$$\text{Gasto\_RS} = (200 \times 0,55) \times 5 \text{ Mbps} = 550 \text{ Mbps}$$

$$\text{Gasto\_SM} = (200 \times 0,3) \times 1 \text{ Mbps} = 60 \text{ Mbps}$$

$$\text{Gasto\_VO} = (200 \times 0,1) \times 10 \text{ Mbps} = 200 \text{ Mbps}$$

$$\text{Gasto\_V} = (200 \times 0,05) \times 5 \text{ Mbps} = 50 \text{ Mbps}$$

$$\text{Gasto\_Total} = 550 + 60 + 200 + 50 = 880 \text{ Mbps}$$

A velocidade mínima encontrada para atender a demanda média de usuários é de 880 Mbps.

A contratação com base na média, em vez do pico, apresenta vantagens financeiras consideráveis. Contratar uma largura de banda que atenda apenas à média economiza recursos, uma vez que não há a necessidade de provisionar para situações extremas que raramente ocorrem. Isso resulta em uma alocação mais eficiente dos recursos financeiros da instituição.

Além disso, a aplicação do modelo LSTM, desenvolvido para predição de usuários na rede discutido nesse trabalho, desempenha um papel fundamental nesse processo. Ao utilizar dados históricos, o LSTM é capaz de prever a quantidade de usuários em diferentes horários, proporcionando uma visão mais precisa das demandas futuras. Essa previsão contribui para a tomada de decisões informada e estratégica na contratação de largura de banda, permitindo que a instituição ajuste seus recursos de acordo com as reais necessidades da comunidade acadêmica.

Portanto, a estratégia de contratação de largura de banda com base na média dos usuários conectados, apoiada pelo modelo LSTM, representa uma abordagem eficaz para otimizar recursos financeiros sem comprometer a qualidade de serviço. Ao utilizar a predição de carga para contratar um link de internet proporcional às necessidades futuras, a instituição evita o pagamento de custos adicionais por ultrapassar a capacidade contratada. Isso resulta em uma gestão mais eficiente dos recursos financeiros, direcionando-os de forma estratégica. Essa prática alinha-se com uma gestão proativa da infraestrutura de rede, garantindo uma conectividade confiável e econômica para toda a comunidade acadêmica.

6 Conclusão e Trabalhos Futuros

6.1 Conclusão

A utilização e a importância da inteligência artificial, principalmente as redes neurais, está aumentando e mostrando resultados promissores em diversas áreas. Em vista disso, o estudo realizado nesse trabalho apresenta como contribuições principais a caracterização, o planejamento e o gerenciamento de recursos relacionados a redes sem fio de larga escala. A pesquisa realizada utiliza dados sólidos, coletados de uma base real da rede institucional da USP e fornecidos por registros para previsão de carga de usuários.

Além disso, o objetivo aqui definido foi de aplicar de técnicas de aprendizado de máquina e o desenvolver um modelo de predição para prever a carga de usuários em cada hora do dia. Para isso foi desenvolvido uma rede neural artificial para previsão de séries temporais com o modelo *Long Short-Term Memory*.

Sendo assim, testes e análises a respeito dos resultados foram desenvolvidas e posteriormente uma análise de possíveis casos de uso como a prevenção de colapso na rede, que ofereceu informações significativas dos pontos críticos e gargalos potenciais, e a economia financeira, ao considerar a média de usuários e aplicar uma contratação de largura de banda mais alinhada com as demandas típicas, demonstrando um uso inteligente dos recursos financeiros da instituição. Essas práticas não apenas economiza custos, evitando provisionamentos excessivos para picos raros, mas também promove uma alocação eficiente dos recursos financeiros, direcionando-os para áreas prioritárias.

Contudo, a predição de carga de usuários nas redes, através de redes neurais, mostra-se uma abordagem viável e promissora. Ao realizar essa análise em uma rede, facilita-se a tomada de decisões em relação a alocação de recursos como a largura de banda, alocação de pontos de acesso, o que impacta diretamente no desempenho, na economia de recursos e economia energética.

6.2 Trabalhos Futuros

Para futuras pesquisas, é recomendável aprofundar os ajustes de hiperparâmetros, buscando não apenas uma configuração ótima, mas também considerando a robustez e generalização do modelo. Essa otimização contínua pode ser explorada em conjunto com algoritmos de ajuste dinâmico de hiperparâmetros, adaptando o modelo às mudanças nas condições da rede ao longo do tempo.

Além disso, uma extensão interessante deste estudo seria a aplicação do modelo de previsão em diferentes instituições conectadas aos servidores da RNP. Isso proporcionaria uma visão mais abrangente e aplicável, permitindo a personalização do modelo de acordo com as características específicas de cada instituição. Essa abordagem contribuiria para a criação de soluções mais adaptáveis e eficientes em contextos diversos.

Outra área promissora para futuras investigações reside na exploração da transferência de aprendizagem. A aplicação desta técnica pode potencializar o desempenho do modelo LSTM no contexto de predição de carga de autenticações em serviços acesso sem fio. Ao pré-treinar o modelo em conjuntos de dados relacionados, como provenientes de outras redes sem fio ou domínios afins, é possível capturar padrões mais amplos e melhorar a capacidade de generalização.

A transferência de aprendizagem oferece a vantagem de aproveitar conhecimentos adquiridos em uma tarefa específica para impulsionar o desempenho em uma tarefa relacionada, mesmo que as condições específicas variem. Essa abordagem pode ser particularmente valiosa ao lidar com características específicas de diferentes instituições conectadas aos servidores da RNP, permitindo uma adaptação mais rápida e eficiente do modelo às nuances individuais de cada contexto.

Em suma, o refinamento contínuo dos hiperparâmetros, a expansão da aplicação do modelo para diversas instituições e a exploração de transferência de aprendizagem representam caminhos promissores para a evolução desta pesquisa, proporcionando resultados mais sólidos e aplicáveis em diferentes cenários de redes sem fio.

Bibliografia

ABINOJA, D.; BEDRUZ, R. A.; JOVELLANOS, K. L.; BANDALA, A. Wireless user estimation using artificial neural networks. In: *2015 International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*. [S.l.: s.n.], 2015. p. 1–5.

AKIBA, T.; SANO, S.; YANASE, T.; OHTA, T.; KOYAMA, M. Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.]: ACM, 2019.

APOSTOLO, G. H.; BERNARDINI, F.; MAGALHãES, L. C. S.; MUCHALUAT-SAADE, D. C. escifi: An energy saving mechanism for wlans based on machine learning. *Energies*, v. 15, n. 2, 2022. ISSN 1996-1073. Disponível em: ⟨https://www.mdpi.com/1996-1073/15/2/462⟩.

BERG, J. *The IEEE 802.11 standardization its history, specifications, implementations, and future*. [S.l.], 2011.

BáSICA, M. da Educação/Secretaria de E. *Portaria nº 33, de 7 de agosto de 2023*. 2023. ⟨https://www.in.gov.br/web/dou/-/portaria-n-33-de-7-de-agosto-de-2023-501491507⟩. Publicado em: 08/08/2023, Edição: 150, Seção: 1, Página: 34.

CHEN, W.; LYU, F.; WU, F.; YANG, P.; REN, J. Flag: Flexible, accurate, and long-time user load prediction in large-scale wifi system using deep rnn. *IEEE Internet of Things Journal*, v. 8, n. 22, p. 16510–16521, 2021.

DING, G.; YUAN, J.; BAO, J.; YU, G. Lstm-based active user number estimation and prediction for cellular systems. *IEEE Wireless Communications Letters*, v. 9, n. 8, p. 1258–1262, 2020.

EDUROAM. *What is eduroam?* 2022. ⟨https://eduroam.org/what-is-eduroam/⟩. Acessado: 09-01-2023.

FANG, L.; XUE, G.; LYU, F.; SHENG, H.; ZOU, F.; LI, M. Intelligent large-scale ap control with remarkable energy saving in campus wifi system. In: *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*. [S.l.: s.n.], 2018. p. 69–76.

FRANK, L. R.; OLIVEIRA, R. M. D.; VIEIRA, A. B.; SILVA, E. F. Improving a smart environment with wireless network user load prediction. In: IEEE. *2021 IEEE Symposium on Computers and Communications (ISCC)*. [S.l.], 2021. p. 1–6.

GUO, J.; CHEN, Y.; ZHU, J.; ZHANG, S. Can we achieve better wireless traffic prediction accuracy? *IEEE Communications Magazine*, v. 59, n. 8, p. 58–63, 2021.

HAYKIN, S. *Redes Neurais: Princípios e prática*. 2.ª. ed. Porto Alegre, RS: Bookman, 2001.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 1997.

HOFSTEDE, R.; ČELEDA, P.; TRAMMELL, B.; DRAGO, I.; SADRE, R.; SPEROTTO, A.; PRAS, A. Flow monitoring explained: From packet capture to data analysis with netflow and ipfix. *IEEE Communications Surveys & Tutorials*, v. 16, n. 4, p. 2037–2064, 2014.

JAIN, A.; MAO, J.; MOHIUDDIN, K. Artificial neural networks: a tutorial. *Computer*, v. 29, n. 3, p. 31–44, 1996.

KIM, T. Y.; OH, K. J.; KIM, C.; DO, J. D. Artificial neural networks for non-stationary time series. *Neurocomputing*, v. 61, p. 439–447, 2004. ISSN 0925-2312. Hybrid Neurocomputing: Selected Papers from the 2nd International Conference on Hybrid Intelligent Systems. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0925231204002371⟩.

KUROSE, J.; ROSS, K. A top-down approach. *Computer Networking*, p. 284, 2013.

MANZOOR, S.; ZHANG, C.; HEI, X.; CHENG, W. Understanding traffic load in software defined wifi networks for healthcare. In: *2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*. [S.l.: s.n.], 2019. p. 1–2.

MEDEIROS, D. S.; NETO, H. N.; LOPEZ, M. A.; MAGALHAES, L. C. S.; SILVA, E. F.; VIEIRA, A. B.; FERNANDES, N. C.; MATTOS, D. M. Análise de dados em redes sem fio de grande porte: Processamento em fluxo em tempo real, tendências e desafios. *Sociedade Brasileira de Computação*, 2019.

NASA Tecnologia. ⟨https://nasatecnologia.com.br/⟩. Accessed on: November 2023.

OPTUNA. *Optuna Documentation*. 2023. ⟨https://optuna.readthedocs.io/en/stable/⟩. Accessed on: November 2023.

PENG, M.; YU, Y.; XIANG, H.; POOR, H. V. Energy-efficient resource allocation optimization for multimedia heterogeneous cloud radio access networks. *IEEE Transactions on Multimedia*, v. 18, n. 5, p. 879–892, 2016.

PINHEIRO, J. M. H.; BECKER, M. *Um estudo sobre algoritmos de Boosting e a otimização de hiperparâmetros utilizando optuna*. 2023.

QIU, C.; ZHANG, Y.; FENG, Z.; ZHANG, P.; CUI, S. Spatio-temporal wireless traffic prediction with recurrent neural network. *IEEE Wireless Communications Letters*, v. 7, n. 4, p. 554–557, 2018.

Rede Nacional de Ensino e Pesquisa. *Quem somos*. 2022. ⟨https://www.rnp.br/sobre⟩. Acessado: 09-01-2023.

RODRIGUES, M. B. d. A.; MENDES, A. C. R.; MENDONCA, M. P. C. de; CARRARA, G. R.; MAGALHAES, L. C. S.; ALBUQUERQUE, C. V. N. de; MEDEIROS, D. S. V.; MATTOS, D. M. F. An efficient strategy with high availability for dynamic provisioning of access points in large-scale wireless networks. In: *2022 5th Conference on Cloud and Internet of Things (CIoT)*. [S.l.: s.n.], 2022. p. 92–99.

SALAM, N.; ABBAS, M. K.; MAHESHWARI, M. K.; CHOWDHRY, B.; NISAR, K. Future mobile technology: Channel access mechanism for lte-laa using deep learning. In: *2021 IEEE 18th Annual Consumer Communications Networking Conference (CCNC)*. [S.l.: s.n.], 2021. p. 1–5.

SUBRAHMANYAM, G.; SATYANARAYANA, P. Neural network based traffic prediction for wireless data networks. *International Journal of Computational Intelligence Systems*, December 2008, p. 379–389, 03 2012.

TAE-EOG; LEE, H.; SREENIVAS, R. S. Token delays and generalized workload balancing for timed event graphs with application to cluster tool operation. In: *2012 20th IEEE International Conference on Network Protocols (ICNP)*. Los Alamitos, CA, USA: IEEE Computer Society, 2006. ISSN 2161-8070. Disponível em: ⟨https://doi.ieeecomputersociety.org/10.1109/COASE.2006.326861⟩.

THANTHARATE, A.; PAROPKARI, R.; WALUNJ, V.; BEARD, C. Deepslice: A deep learning approach towards an efficient and reliable network slicing in 5g networks. In: IEEE. *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. [S.l.], 2019. p. 0762–0767.

WANG, J.; JIANG, C. Machine learning paradigms in wireless network association. In: _____. [S.l.: s.n.], 2018. p. 1–9.

YOU, C.; CHANDRA, K. Time series models for internet data traffic. In: *Proceedings 24th Conference on Local Computer Networks. LCN'99*. [S.l.: s.n.], 1999. p. 164–171.