UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CURSO DE GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

João Jandre Paraquetti

Predição de Ocupação de Leitos Hospitalares de Terapia Intensiva no Curto Prazo Utilizando Redes Neurais Artificiais

Florianópolis
2023

João Jandre Paraquetti

Predição de Ocupação de Leitos Hospitalares de Terapia Intensiva no Curto Prazo Utilizando Redes Neurais Artificiais

Trabalho de Conclusão de Curso do Curso de Graduação em Ciências da Computação do DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Ciências da Computação.
Orientadora: Profa. Dra. Jerusa Marchi

Florianópolis
2023

RESUMO

Prever com eficácia a demanda por atendimentos é importante em praticamente todos os tipos de serviços ofertados ao público, pois alocar menos recursos do que o necessário pode levar a perda da qualidade do atendimento ao consumidor e alocar recursos a mais pode significar desperdício de tais recursos. No caso específico de hospitais, a espera demasiada por atendimento pode levar ao óbito de pacientes e a alta demanda por serviços hospitalares públicos obriga um uso ótimo de recursos, evitando desperdício. Este trabalho dá uma breve visão sobre o problema de predição de ocupação de leitos hospitalares, e usando os dados disponíveis pelo DATASUS dos hospitais do estado de Santa Catarina, faz uma decomposição da série temporal da ocupação de leitos, buscando explicar comportamento das mesmas, os dados obtidos e tratados foram usados para treinamento de diversas redes LSTM, com o objetivo de prever a ocupação de leitos de UTI no estado de Santa Catarina. Finalmente, o trabalho faz uma comparação entre os modelos desenvolvidos, concluindo que a rede que apresentou os melhores resultados foram obtidos com o modelo do terceiro governo do período estudado, utilizando apenas 4 neurônios na camada oculta, e considerando 3 semanas anteriores para a predição.

Palavras-chave: Redes Neurais Artificiais. Predição. Leitos Hospitalares.

ABSTRACT

Effectively predicting demand for healthcare services is important in virtually all types of public services, as allocating fewer resources than necessary can result in a loss of quality in consumer care, while allocating excessive resources can lead to wastage. In the specific case of hospitals, excessive waiting times for care can lead to patient mortality, and the high demand for public hospital services necessitates optimal resource utilization to avoid waste. This study provides a brief overview of the problem of predicting hospital bed occupancy and, using the available data from DATASUS on hospitals in the state of Santa Catarina, decomposes the time series of bed occupancy to explain its behavior. The obtained and processed data were used to train various LSTM networks with the aim of predicting ICU bed occupancy in the state of Santa Catarina. Finally, the study compares the developed models, concluding that the best results were achieved with the model from the third government of the period studied, using only 4 neurons in the hidden layer and considering the previous 3 weeks for prediction.

Keywords: Artificial Neural Networks. Prediction. Hospital beds.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIH	Autorização de Internação Hospitalar
LSTM	*Long short term memory*
MAPE	Erro Percentual Absoluto Médio
RMSE	Raiz quadrada do erro médio
RNA	Rede neural artificial
SIH	Sistema de Internações Hospitalares
UTI	Unidade de tratamento intensivo

SUMÁRIO

1 INTRODUÇÃO

Projetar e implementar políticas eficazes de gestão de capacidade hospitalar e decisões de alocação de profissionais é um desafio crítico em todos os sistemas de saúde. Um desencontro entre número de leitos disponíveis e a demanda, assim como a quantidade de profissionais de saúde alocados podem afetar negativamente diversos indicadores de performance dos hospitais, como tempo de espera, tamanho da fila de espera, qualidade do atendimento, assim como a satisfação dos pacientes e profissionais de saúde (TELLO *et al.*, 2022).

Levando isso em conta, existe uma necessidade amplamente reconhecida de prever a ocupação de leitos de hospital. Quanto melhores previsões pudermos fazer, mais eficientemente poderemos planejar com antecedência e, como resultado, o uso de recursos é otimizado e melhores cuidados podem ser prestados aos pacientes (KUTAFINA *et al.*, 2019).

O problema de previsão de ocupação de leitos já é explorado há décadas, com diversas soluções propostas. Classicamente, as abordagens numéricas eram populares, diversos autores propunham o uso de regressões, como em (BEENHAKKER, 1963), muitas vezes utilizando-se de séries temporais, como proposto em (LITTIG; ISKEN, 2007). O modelo auto-regressivo integrado de médias móveis, ou ARIMA, também foi outra técnica explorada por autores, como em (EARNEST *et al.*, 2005) onde foi feito um estudo retrospectivo utilizando esse modelo para prever o número de leitos ocupados durante o surto de SARS de 2003 em Singapura.

Nas últimas décadas, abordagens orientadas a dados e aprendizado de máquina provaram sua eficiência para tarefas de previsão. No entanto, apenas um progresso limitado foi feito na aplicação desses métodos à previsão de ocupação de leitos hospitalares (KUTAFINA *et al.*, 2019). Entretanto, nos últimos anos, houve um crescimento do uso de técnicas de aprendizado de máquina para previsão de leitos de UTI, especialmente durante a pandemia de COVID-19.

Em (TELLO *et al.*, 2022), os autores utilizam *k-means clustering* junto de uma máquina de vetores de suporte para predizer a ocupação de leitos de um hospital na Pensilvânia, já (SCHIELE; KOPERNA; BRUNNER, 2021) propuseram um modelo de redes neurais recursivas para previsão de ocupação de leitos de UTI no hospital universitário de Augsburgo. No contexto da pandemia, (BRAGA *et al.*, 2021) utilizaram uma rede neural artificial para predizer a ocupação de leitos durante a pandemia de COVID-19 no estado do Pará.

No Brasil, todos os dados coletados por hospitais públicos são armazenados pelo ministério da saúde (CARVALHO, 2009), de forma que, técnicas capazes de processar grandes quantidades de dados poderiam usar essas informações para construir modelos precisos, como é o caso das técnicas de aprendizado de máquina. Este trabalho visa aplicar

redes neurais do tipo Long Short Term Memory para previsão de ocupação de leitos hospitalares em curto prazo.

1.1 OBJETIVO GERAL

Fazer um estudo preliminar da aplicabilidade de redes neurais do tipo Long Short Term Memory para a predição de leitos hospitalares considerando demandas de curto prazo.

1.2 OBJETIVOS ESPECÍFICOS

- Compreender quais fatores influenciam o aumento da demanda por leitos em situações não críticas.
- Arquitetar e treinar modelos de rede neurais artificiais do tipo *Long short term memory* (LSTM), baseado-se nos dados fornecidos pelo DATASUS sobre a ocupação de leitos no estado de Santa Catarina, com o objetivo de predição de leitos hospitalares em curto prazo.
- Fazer uma comparação entre as redes LSTM desenvolvidas neste trabalho, visando encontrar o melhor modelo. Para tal serão aplicados critérios quantitativos e qualitativos como acurácia da predição e necessidade de dados específicos para treinamento.

1.3 ORGANIZAÇÃO DO TRABALHO

O restante do trabalho está disposto da seguinte maneira: o capítulo 2 apresenta a fundamentação teórica necessária para o entendimento do trabalho, explicando sobre o SIH/SUS, séries temporais e redes neurais. O capítulo 3 expõe alguns trabalhos correlacionados à este. O capítulo 4 possui o desenvolvimento do trabalho, passando desde uma explicação sobre o tratamento de dados, até o treinamento das redes desenvolvidas. Finalmente, o capítulo 5 apresenta uma conclusão do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresentará o Sistema de Informações Hospitalares do SUS, assim como os conceitos básicos de séries temporais e redes neurais artificiais, além de uma visão geral dos diversos tipos de RNAs e respectivos usos mais populares. Ademais, as redes neurais *feedforward* e as LSTM serão apresentadas com maior detalhes.

2.1 SISTEMA DE INFORMAÇÕES HOSPITALARES-SIH

O Sistema de Informações Hospitalares, ou Sistema de Internações Hospitalares (SIH), teve seu início na década de 1970, foi implantado com a intenção de controlar o pagamento dos serviços prestados pelos hospitais contratados (LESSA *et al.*, 2000). Até 1991, o sistema passou por diversos nomes, para então finalmente ser renomeado para Sistema de Informações Hospitalares (LEVCOVITZ; PEREIRA, 1993). Todo o acervo de informações e valores do sistema antigo passou a compor a base do SIH/SUS(LESSA *et al.*, 2000).

O SIH passou por várias plataformas em mainframes UNISYS e ABC-BULL, na fase de processamento centralizado. Foi o primeiro sistema do DATASUS[1] a ter captação implementada em microcomputadores e descentralizada nos próprios usuários, encerrando a era dos polos de digitação. O processamento das Autorizações de Internação Hospitalar, ou Autorização de Internação Hospitalar (AIH), continuou centralizado até ser descentralizado para os gestores de Secretaria de Saúde em abril de 2006, usando plataforma Windows, SGBD Firebird e Linguagem de programação Delphi, que é o estado em que se encontra atualmente(IBGE, 2023).

O sistema transcreve todos os atendimentos provenientes de internações hospitalares que foram financiadas pelo SUS, e após o processamento, gera relatórios para os gestores, possibilitando fazer os pagamentos dos estabelecimentos de saúde. Além disso, o nível federal recebe mensalmente uma base de dados de todas as internações autorizadas (aprovadas ou não para pagamento), para que possam ser repassados às Secretarias de Saúde (IBGE, 2023).

A AIH é o documento hábil para identificar o paciente e os serviços prestados sob regime de internação hospitalar pelo SUS. Fornece informações para o gerenciamento do sistema e, através dele, os hospitais, profissionais e serviços auxiliares de diagnóstico e terapia receberão pelos serviços prestados ao usuário (LESSA *et al.*, 2000).

A AIH pode ser separada em 5 seções: identificação do hospital, caracterização da internação, procedimentos especiais, serviços profissionais e caracterização da assistência prestada. Para os objetivos do trabalho, apenas as seções de caracterização da internação e procedimentos especiais serão relevantes, pois é onde estão descritas as informações sobre o paciente e informações sobre internações sobre o uso da unidade de tratamento

[1] Departamento de Informática do SUS

intensivo, Unidade de tratamento intensivo (UTI). Uma descrição completa dos dados das AIH pode ser encontrada no anexo 1 de (LESSA *et al.*, 2000).

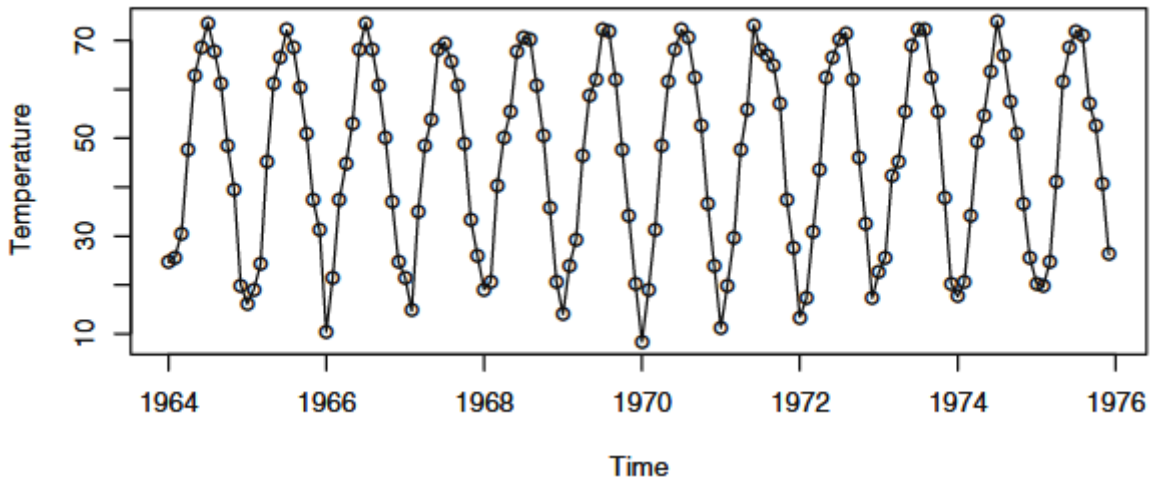
Os dados disponibilizados livremente pelo DATASUS são um conjunto reduzido e anonimizado das AIH, de forma que não é possível identificar os pacientes de nenhuma forma, o que poderia ser um fator limitante para certos tipos de estudos. Entretanto como este trabalho visa apenas estudar o comportamento da ocupação de leitos de UTI, não é necessária a identificação de pacientes.

2.2 SÉRIES TEMPORAIS

Uma série temporal é muitas vezes o resultado da observação de algum processo, onde valores são coletados a partir de medições feitas em instantes de tempo uniformemente espaçados, de acordo com uma determinada taxa de amostragem. Uma série temporal pode, portanto, ser definida como um conjunto de instantes de tempo contíguos (ESLING; AGON, 2012).

Inúmeras áreas de conhecimento fazem esse tipo de coleta de dados. Na meteorologia, são observadas as temperaturas diárias e precipitação anual. Na agricultura, são registrados os números anuais de produção agrícola e pecuária, erosão do solo e exportações. Nas empresas, são observadas as taxas de juros semanais, preços de ações de fechamento diários, índices de preços mensais, números de vendas anuais entre outros. A quantidade de exemplos desse tipo de coleta de dados é enorme (CRYER, 1986).

Figura 1 – Temperaturas médias mensais, Dubuque, Iowa



Fonte: (BROCKWELL; DAVIS, 2002)

Uma série temporal discreta é aquela em que o conjunto T_0 de tempos nos quais as observações são feitas é um conjunto discreto, como por exemplo na figura 1, quando as

observações são feitas em pontos fixos [2]. Séries temporais contínuas são obtidas quando as observações são registradas continuamente durante algum intervalo de tempo, por exemplo, quando $T_0 = [0, 1]$ (BROCKWELL; DAVIS, 2002).

Existem algumas características comuns que são observadas ao trabalhar com séries temporais, como tendência, sazonalidade, ciclos e estacionariedade. Uma tendência existe quando há um aumento ou diminuição de longo prazo nos dados, não necessariamente linear (HYNDMAN; ATHANASOPOULOS, 2018). Um padrão sazonal ocorre quando uma série temporal é afetada por fatores sazonais, como a época do ano ou o dia da semana. A sazonalidade é sempre de um período fixo e conhecido (HYNDMAN; ATHANASOPOULOS, 2018). Um ciclo ocorre quando os dados exibem aumentos e quedas que não são de uma frequência fixa. A duração dessas flutuações é geralmente de pelo menos 2 anos (HYNDMAN; ATHANASOPOULOS, 2018). Uma série temporal é dita estacionaria se não possui uma mudança sistemática na média, ou seja, sem tendência, se não possui nenhuma mudança sistemática na variância, e se variações estritamente periódicas tenham sido removidas (CHATFIELD, 2013).

Um guia importante para as propriedades de uma série temporal é dado pelos coeficientes de autocorrelação da amostra, que medem a correlação entre observações a diferentes distâncias no tempo, comumente chamados de atrasos, ou *lags*. Esses coeficientes geralmente fornecem informações sobre o modelo de probabilidade que gerou os dados(CHATFIELD, 2013). Quando os dados têm uma tendência, as autocorrelações para pequenas distancias no tempo tendem a ser grandes e positivas, porque as observações próximas no tempo também têm valores próximos. Quando os dados são sazonais, as autocorrelações serão maiores para os *lags* sazonais (atrasos de tempo múltiplos do período sazonal) do que para outros *lags*, isso ocorre porque os dados exibem um padrão repetitivo em intervalos regulares. Séries temporais que não mostram nenhuma autocorrelação são chamadas de ruído branco. (HYNDMAN; ATHANASOPOULOS, 2018).

Uma série pode ser univariada, ou seja, quando uma única variável é observada ao longo do tempo; ou multivariada, quando várias séries abrangem simultaneamente várias dimensões dentro do mesmo intervalo de tempo (ESLING; AGON, 2012). Uma série temporal pode cobrir todo o conjunto de dados fornecidos pela observação de um processo, e pode ter um comprimento considerável. Além disso, elas são consideradas suaves, isto é, valores subsequentes estão dentro de faixas predizíveis uns dos outros (ESLING; AGON, 2012).

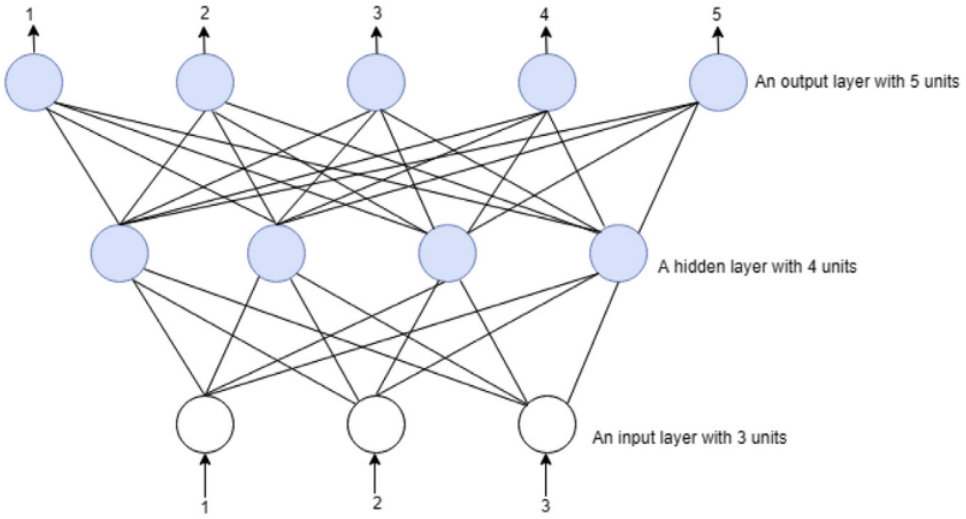
Geralmente existem dois propósitos para fazer uma análise de séries temporais: entender ou modelar o mecanismo estocástico que gera uma série observada, e predizer os valores futuros da série baseando-se no histórico daquela série e possivelmente, outras séries relacionadas (CRYER, 1986).

[2] Neste gráfico, as observações são representadas pelos círculos no gráfico, enquanto que a curva traçada é apenas uma aproximação dos dados da série seriam caso a mesma fosse contínua

2.3 REDES NEURAIS ARTIFICIAIS

A computação clássica é baseada num conjunto explícito de instruções programadas, que datam do trabalho de Babbage, Turing e von Neumann; redes neurais artificiais representam um paradigma computacional alternativo, onde a solução para um problema é aprendida a partir de um conjunto de exemplos (BISHOP, 1994). Redes neurais artificiais, ou Rede neural artificial (RNA), são um subgrupo do aprendizado de máquina, seu nome e estrutura são baseados nas redes de neurônios que constituem os cérebros dos seres vivos munidos de sistema nervoso. Uma RNA é composta por um conjunto de neurônios, ou nodos, que operam sobre as informações recebidas e se comunicam com outros neurônios (SVOZIL; KVASNICKA; POSPICHAL, 1997).

Figura 2 – Rede neural *feedforward* com 3 camadas



Fonte: (ABIODUN *et al.*, 2018)

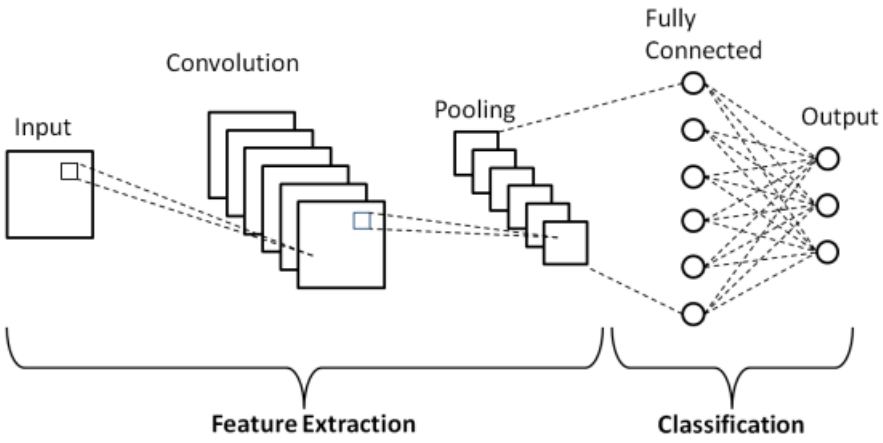
A figura 2 apresenta a estrutura usual de uma Rede Neural do tipo *feedforward*, onde os nodos representam os neurônios da rede, além disso, os arestas representam as conexões sinápticas entre os neurônios, geralmente nesse tipo de rede os neurônios são organizados em camadas de entrada, de saída e camadas ocultas, nesse caso, a rede possui três neurônios na camada de entrada, quatro neurônios na camada oculta e cinco neurônios na camada de saída.

Existem dois tipos principais de processos de aprendizado em RNAs: aprendizado supervisionado e aprendizado não supervisionado. No aprendizado supervisionado, a rede sabe a saída desejada e os pesos dos neurônios são ajustados de forma que a saída desejada e a saída obtida sejam o mais próximas possível. Já no aprendizado não supervisionado, a saída desejada não é sabida, porém são providos um grupo de fatos para a rede, e a mesma deve alcançar um estado estável dentro de um certo número de iterações (SVOZIL; KVAS-

NICKA; POSPICHAL, 1997). Um exemplo de RNA com aprendizado não supervisionado são as redes de Konohen (KOHONEN, 2012).

Dentro do grupo das RNAs com aprendizado supervisionado existe uma multitude de tipos de RNAs, como por exemplo, redes neurais profundas, que ficaram populares na literatura pois são capazes de lidar com enormes quantidades de dados (ALBAWI; MOHAMMED; AL-ZAWI, 2017), dentro das redes profundas, ainda existem as redes neurais convolucionais, utilizadas geralmente para problemas de reconhecimento de imagem, visão computacional e processamento de linguagem natural (ALBAWI; MOHAMMED; AL-ZAWI, 2017).

Figura 3 – Rede neural convolucional



Fonte: (PHUNG; RHEE, 2019)

A figura 3 ilustra um diagrama esquemático da arquitetura de uma rede neural convolucional para reconhecimento de imagens, que consiste em cinco camadas diferentes: entrada, convolução, agrupamento, totalmente conectada e saída. A camada de entrada especifica um tamanho fixo para as imagens de entrada, ou seja, as imagens podem ter que ser redimensionadas para passarem pela rede. A imagem é então convolucionada com vários *kernels* usando pesos compartilhados. Em seguida, as camadas de agrupamento reduzem o tamanho da imagem enquanto tentam manter as informações nela contidas. Essas duas camadas compreendem a parte de extração de características. Em seguida, as características extraídas são ponderadas e combinadas na camada totalmente conectada. Isso representa a parte de classificação da rede neural convolucional. Finalmente, há um neurônio de saída para cada categoria de objeto na camada de saída (PHUNG, V. H.; RHEE, E. J., 2018).

As RNAs geralmente são treinadas em épocas, ou seja, cada vez que a rede processar todo o conjunto de dados, uma época se passou. Então, os pesos dos neurônios são reajustados. Porém, é muito comum que seja definido um tamanho de *batch*, que vai definir o número de amostras que serão propagadas na rede antes que os pesos sejam

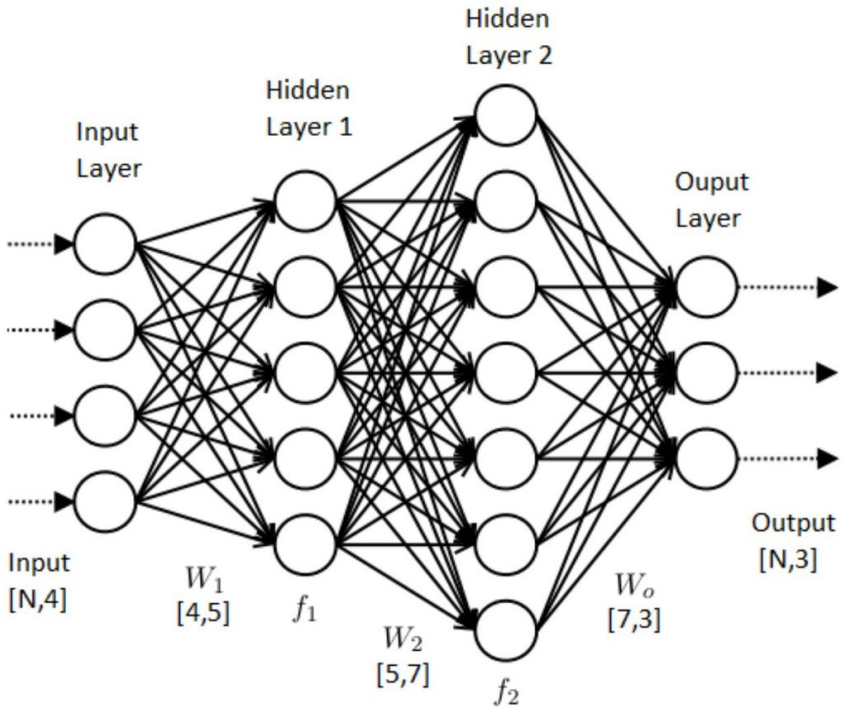
reajustados, nesse caso, a cada época, a rede "aprende" diversas vezes. Existe também um tipo especial de aprendizado, onde o tamanho do *batch* é 1, ou seja, os pesos da rede são mudados a cada nova entrada, nesse caso, isso é chamado de aprendizado *online* (ALPAYDIN, 2020).

2.3.1 Redes Neurais Artificiais Feedforward

Um tipo comum de rede neural com aprendizado supervisionado são as redes do tipo *feedforward*, que é composta por conjuntos de neurônios, onde cada um processa a informação recebida em sua entrada, utilizando alguma função não-linear, como por exemplo uma função sigmoide, aplicada na soma ponderada de todos os valores da entrada do mesmo, então esse valor de saída é passado para os neurônios conectados mais à frente.

Nesse tipo de rede, as informações sempre são passadas apenas para os neurônios seguintes, sem retroalimentação ou ciclos, o que as diferencia das redes chamadas redes neurais recursivas. Assim, os neurônios são organizados em camadas, com três tipos de camada, uma camada de entrada, uma ou mais camadas escondidas, e uma camada de saída. Cada neurônio recebe informações de todos os neurônios da camada anterior a sua.

Figura 4 – Rede neural artificial



Fonte: Data Science Central. Disponível em: <https://www.datasciencecentral.com/the-artificial-neural-networks-handbook-part-1/>. Acesso em: 22 nov. 2022.

Um exemplo detalhado de rede neural *feedforward* é apresentado na figura 4. Nesse exemplo rede está organizada em quatro camadas, sendo uma delas a camada de entrada,

com quatro neurônios, uma camada de saída, com três neurônios, e duas camadas ocultas, com cinco e sete neurônios, respectivamente. Também é possível visualizar que a rede é completamente conectada, já que cada neurônio de uma camada está conectado a todos os outros da camada mais à diante. Além disso, são ilustradas na figura o tamanho das matrizes de pesos de cada camada, como por exemplo a matriz W_1, que possui 4 linhas e 5 colunas, é nessas matrizes que ficam armazenados os pesos de cada conexão entre os neurônios, ou seja, é onde o que a rede aprendeu fica armazenado.

A desvantagem de se utilizar redes com várias camadas é que o aprendizado se torna mais difícil, a solução comum para este problema é a utilização do algoritmo de *error back-propagation*, ou retropropagação do erro, no aprendizado. Inicialmente, os pesos das entradas dos neurônios são aleatórios, então, a cada ciclo de propagação/adaptação, o erro é calculado com base nas saídas obtidas contra as saídas desejadas, e esse sinal de erro é então retro-propagado da camada de saída para cada elemento da camada anterior que contribuiu diretamente para a formação da saída (ROISENBERG, s.d.).

Cada elemento da camada anterior recebe apenas uma porção do sinal de erro total, proporcional apenas à contribuição relativa de cada elemento na formação da saída. Esse processo se repete para cada camada até que cada elemento da rede receba um sinal de erro proporcional a sua contribuição para o erro total. Baseado nesse sinal, os pesos das conexões são recalculados.

2.3.2 Redes Neurais Long Short Term Memory

Ao contrário das redes *feedforward*, as redes *Long short term memory*, ou LSTM, são do tipo recorrentes, ou seja, possuem retroalimentação, o que cria possibilidade da rede ter uma memória de curto prazo. Entretanto, aprender a armazenar informações em intervalos de tempo prolongados por meio de retropropagação recorrente leva muito tempo, principalmente devido ao fluxo de retorno de erro insuficiente e decadente (HOCHREITER; SCHMIDHUBER, 1997). Com métodos convencionais de otimizar redes neurais recorrentes, como *Real time recurrent learning* (ROBINSON; FALLSIDE, 1987) e *Back-propagation through time* (WILLIAMS; ZIPSER, 1995) sinais de erro "retrocedendo no tempo" tendem a explodir, ou sumir, com o primeiro caso levando a pesos oscilantes, e o segundo levando a tempos de treinamento muito longos ou simplesmente não funcionando (HOCHREITER; SCHMIDHUBER, 1997).

Tentando resolver este problema, é que as redes neurais LSTM foram propostas em 1997 por Sepp Hochreiter e Jurgen Schmidhuber (HOCHREITER; SCHMIDHUBER, 1997). A proposta dos mesmos era uma nova arquitetura de rede neural, em conjunto com um algoritmo de aprendizado baseado em gradiente. E segundo (HOCHREITER; SCHMIDHUBER, 1997) "Ela consegue aprender a transpor intervalos de tempo superiores a 1000 passos, mesmo tendo sequências de entrada ruidosas ou incompressíveis, sem perda da capacidade com tempos curtos de atraso".

O tipo de processo de aprendizado utilizado com redes LSTM é o aprendizado supervisionado, explicado na seção 2.3. Durante o processo de aprendizado, as entradas são transformadas e adicionadas à memoria de curto prazo. Então o estado da memória de curto prazo é usado para determinar qual parte da memória de longo prazo não é mais necessária, abrindo espaço para novas informações na memória de longo prazo. A memória de curto prazo é então adicionada à memoria de longo prazo, atualizando a mesma. Finalmente, a memória de longo prazo é utilizada para modificar a memória de curto prazo (YE, 2023).

Redes LSTM são aplicadas em uma multitude de áreas, como por exemplo para reconhecimento de fala (SAK; SENIOR; BEAUFAYS, 2014), controle de robôs (MAYER *et al.*, 2008), e na área da saúde (SCHMIDHUBER; BLOG, 2020). Esse tipo de rede é especialmente efetiva para fazer predições baseadas em séries temporais, já que ela consegue capturar atrasos de grande duração entre os eventos. Ao contrário das redes neurais convencionais, que são compostas por neurônios ou nodos, redes LSTM geralmente são compostas por células, que possuem um estado interno[3], portões de entrada, saída e de esquecimento. Um exemplo de uma unidade deste tipo é mostrada na figura 5.

A função do *forget gate* é decidir quais informações não são mais úteis no estado da célula, ele é alimentado com a saída anterior da célula e a entrada atual, as entradas são multiplicadas por uma matriz de pesos e então adicionadas de um viés (ACADEMY, 2023). O resultado é passado por uma função de ativação que fornece uma saída binária, geralmente a sigmoide logística (GRAVES, 2012). Se para um determinado estado de célula a saída for 0, a informação é esquecida, se a saída for 1, a informação é retida para uso futuro.

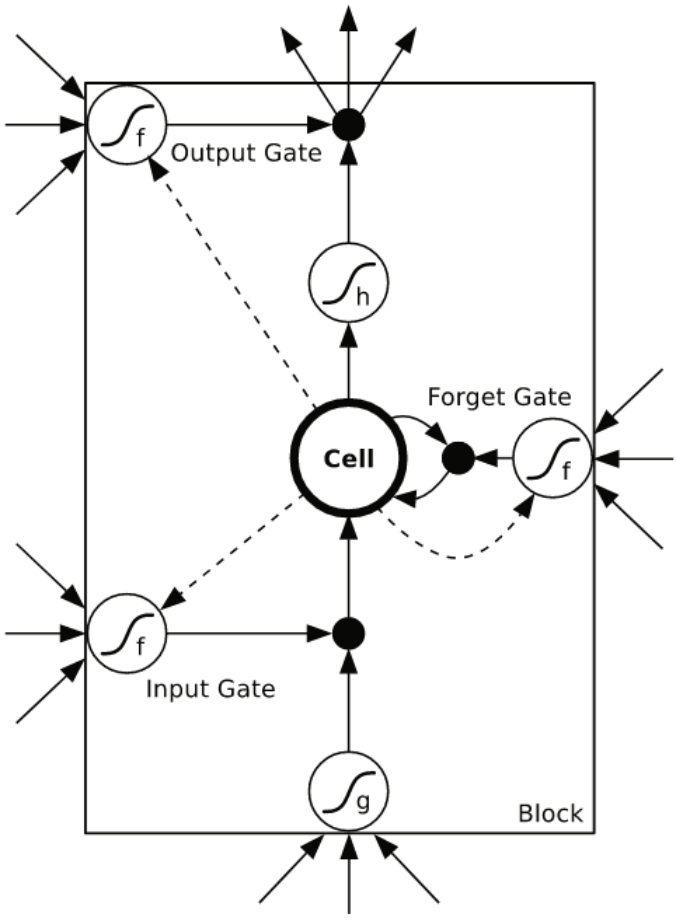
A regulagem de quais novas informações serão adicionadas ao estado da célula é feita através do *input gate*. Similarmente ao *forget gate*, o *input gate* recebe a saída anterior da célula e a entrada atual, multiplica as mesmas por uma matriz de pesos, adicionando um viés e finalmente passando por uma função de ativação que fornece uma saída binária, como a sigmoide logística. Essa saída então é multiplicada pela saída da função f_g, fazendo com que o resultado da mesma seja adicionado ou não ao estado interno da célula (GRAVES, 2012). O *output gate* segue a mesma lógica do *input gate*, porém com o objetivo de decidir o quanto o estado interno da célula vai influenciar na saída.

Nenhuma função de ativação é aplicada dentro da célula. A função de ativação f_f dos portões é geralmente a sigmoide logística, de modo que as ativações dos portões estejam entre 0 (portão fechado) e 1 (portão aberto). As funções de ativação de entrada e saída da célula (f_g e f_h) são geralmente a tangente hiperbólica ou a sigmoide logística, embora em alguns casos f_h seja a função de identidade (GRAVES, 2012).

As linhas tracejadas representam conexões internas, que permitem que os portões da célula tenham acesso ao estado interno da célula, mesmo quando o portão de saída está

[3] Também conhecido como CEC ou *Constant Error Carousel*

Figura 5 – Arquitetura de uma célula LSTM

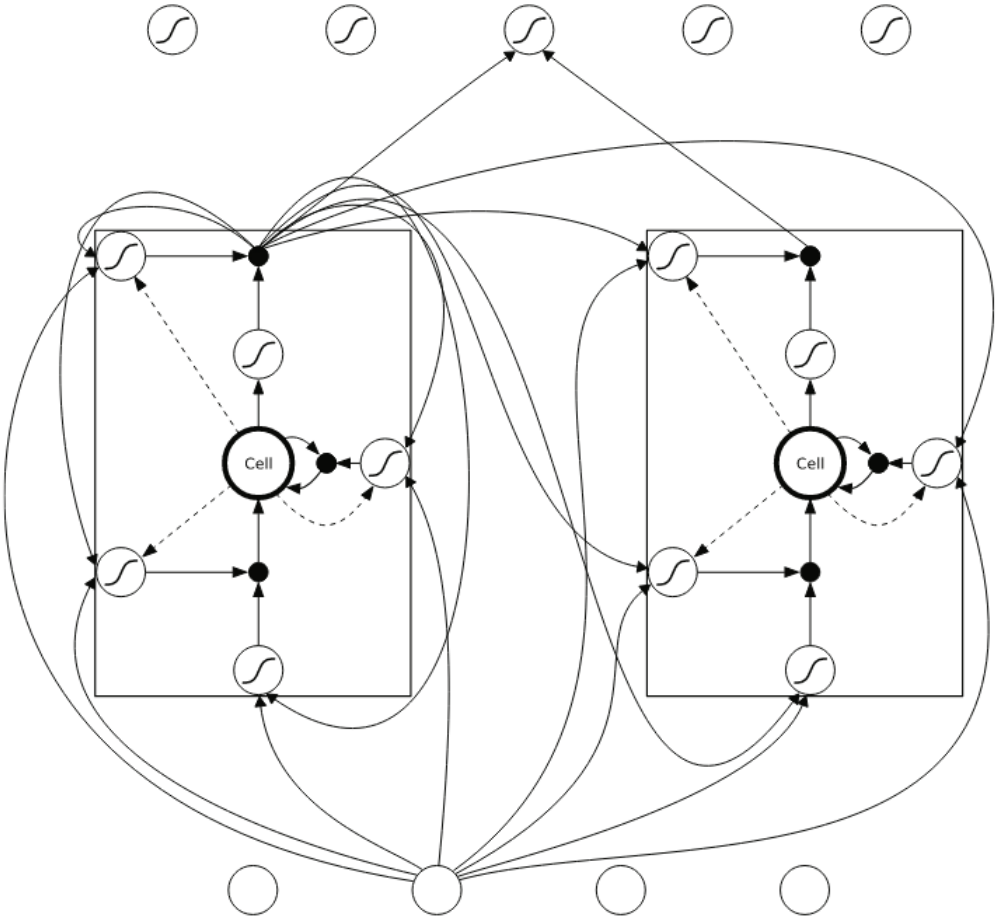


Fonte: (GRAVES, 2012)

fechado, essas conexões são chamadas de *peepholes*. Os *peepholes* não estavam presentes na arquitetura original das LSTMs, tendo sido propostos para melhor a performance em tarefas altamente não lineares, permitindo que a célula aprenda a medir intervalos de tempo precisos (GERS; SCHRAUDOLPH; SCHMIDHUBER, 2002). Nenhuma das outras conexões dentro do bloco são ponderadas. As únicas saídas do bloco para o resto da rede emanam da multiplicação da porta de saída, ou seja, de certo modo, cada célula se comporta como um perceptron, recebendo n entradas porém emanando apenas uma saída.

Um exemplo de rede LSTM pode ser observado na figura 6, a rede demonstrada possui quatro neurônios de entrada, uma camada escondida com duas células LSTM e cinco neurônios de saída. Para não poluir a imagem, nem todas as conexões são mostradas. É importante notar que cada célula tem quatro entradas, porém apenas uma saída, também é interessante notar as conexões recorrentes das células.

Figura 6 – Rede LSTM



Fonte: (GRAVES, 2012)

3 TRABALHOS CORRELATOS

Existem diversas propostas, descritas na literatura, de algoritmos de predição de leitos hospitalares, sendo que a maioria das soluções mais atuais focam na utilização do aprendizado de máquina para resolver essa tarefa. Foi feita uma revisão sistemática da literatura existente, com objetivo de encontrar os trabalhos mais relevantes sobre predição de ocupação de leitos de UTI utilizando aprendizado de máquina.

Foram pesquisados artigos escritos na língua inglesa sobre predição de ocupação de leitos de UTI em curto prazo com aprendizado de máquina, focando no uso de redes neurais artificiais. As pesquisas foram feitas no Google Scholar, por se tratar de uma ferramenta amplamente disponível e que conecta diversas bibliotecas digitais.

Tabela 2 – Revisão bibliográfica.

Termos pesquisados	**Número de resultados obtidos**
prediction “ICU beds” “machine learning” “short term”	474
prediction “ICU beds” “neural network” “short term”	234
prediction “intensive care unit beds” “neural network”	62
prediction “intensive care unit beds” “neural network” “short term”	28

Fonte: Elaboração própria a partir de buscas no Google Scholar.

Na tabela Tabela 2 estão dispostos os termos pesquisados e a quantidade de resultados obtidos. A partir dos resultados obtidos ordenados por ordem de relevância, foram selecionados 45 artigos à partir dos seus títulos. Da leitura do resumo e da conclusão, os 10 artigos mais correlacionados à este foram selecionados. Esses foram lidos integralmente, e 3 artigos foram selecionados para serem descritos nesta seção, por detalharem o problema explorado, explicarem aprofundadamente a solução, e apresentarem os resultados de forma clara. Poucos trabalhos fortemente correlatos foram encontrados.

3.1 PREDICTING INTENSIVE CARE UNIT BED OCCUPANCY FOR INTEGRATED OPERATING ROOM SCHEDULING VIA NEURAL NETWORKS

(SCHIELE; KOPERNA; BRUNNER, 2021) exploram o problema de *master surgery scheduling*, onde tenta-se agendar cirurgias de forma a alocar as especialidades médicas para as diferentes salas cirúrgicas disponíveis, para que as cirurgias sejam realizadas com a maior eficiência possível (BOVIM *et al.*, 2020). Este trabalho se mostra importante pois cria uma ferramenta poderosa para que administradores possam testar decisões de configurações de salas de operações para cada especialidade, para que possam melhor administrar os recursos disponíveis. Este trabalho correlato foi escolhido por fazer uma modelagem aprofundada e bem explicada do problema da ocupação de leitos, entretanto, os mesmos fazem uso de uma base de dados muito mais precisa do que a disponível para o

desenvolvimento do presente trabalho, o que impossibilita uma abordagem tão minuciosa. Além disso, este trabalho visa utilizar-se de redes LSTM para a previsão de leitos, ao contrário de (SCHIELE; KOPERNA; BRUNNER, 2021).

Utilizando-se dos registros de cirurgias, do departamento de emergência, da enfermaria, da unidade de cuidados intermediários e da UTI, os autores propõem o uso de uma rede neural artificial para predizer o número de leitos ocupados na UTI em certo dia, levando até D dias em conta. Para isso, foi criado um modelo levando em conta os possíveis caminhos que um paciente pode fazer dentro de um hospital, como por exemplo ser admitido na enfermaria, ir para a sala de operações, de volta pra enfermaria, para então ter alta. Além disso, foi levado em conta o tempo de estadia do paciente em cada ala do hospital.

Foram testadas algumas arquiteturas de redes neurais no estudo, porém a arquitetura que se saiu melhor tinha duas camadas escondidas, uma com 200 neurônios e a segunda com 50, a função de ativação escolhida foi a ReLu, cada modelo foi treinado em até 100000 épocas, utilizando o gradiente descendente estocástico (AMARI, 1993) com uma taxa de aprendizado constante de 0,00001. Com essa arquitetura e todos os dados disponíveis no estudo, os autores obtiveram um erro quadrático médio de 3.46.

3.2 COVID-19 ICU DEMAND FORECASTING: A TWO-STAGE PROPHET-LSTM APPROACH

No contexto da pandemia de COVID-19, (BORGES; NASCIMENTO, 2022) propuseram um modelo para predizer o número de entradas de pacientes nas UTIs da cidade de São José dos Campos, em São Paulo. Este modelo se mostra interessante por integrar duas técnicas de aprendizado de máquina: o Prophet, um modelo de predição recentemente introduzido pelo Facebook (TAYLOR; LETHAM, 2018) e *Long-Short Term Memory Network*, ou LSTM, que é um dos modelos mais populares para predição de séries temporais (SEZER; GUDELEK; OZBAYOGLU, 2020). Este trabalho correlato foi escolhido por aplicar LSTMs ao problema da predição da ocupação de leitos de UTI, entretanto, o mesmo faz uso conjunto do Prophet, que não será utilizado para este trabalho, já que o mesmo não utiliza *lags* com tamanho grande, que são o principal motivo de (BORGES; NASCIMENTO, 2022) terem utilizado o Prophet, além disso este trabalho estuda apenas a ocupação de leitos de UTI fora do período pandêmico, ao contrário de (BORGES; NASCIMENTO, 2022).

Neste trabalho, os autores utilizaram o número de casos de COVID-19 diários, o índice de isolação social, a ocupação de leitos hospitalares regional, todos esses dados referentes à cidade em estudo, além disso, os mesmos utilizaram informações do Plano-SP, que foi o plano de reabertura do estado de São Paulo, e finalmente uma última variável, criada pelos próprios autores e denominada “Vax”, que diz respeito a um índice artificial criado para refletir a porcentagem da população que provavelmente precisaria de um leito

de UTI.

A arquitetura proposta utilizava o Prophet como primeiro estágio de predição, com o mesmo lidando com as variáveis que possuíam atraso temporal em relação a predição, o LSTM foi utilizado como segundo estágio da predição, que recebia como entrada a saída do Prophet, junto com as variáveis que não possuíam atraso temporal.

O LSTM utilizado pelos autores foi configurado manualmente, e possuía um *batch size* de 5, com uma camada de entrada de 50 neurônios, duas camadas escondidas, uma com 50 e outra com 25 neurônios, a função de ativação escolhida foi a LeakyReLU, com uma taxa de *dropout* de 0,1. Além disso, a função de perda utilizada foi o erro quadrático médio, o otimizador escolhido foi o Adam, e a métrica de avaliação foi a raiz quadrada do erro médio, finalmente, os modelos foram treinados por 200 épocas, utilizando parada antecipada. Utilizando essa arquitetura, os autores atingiram uma média de erros médios quadráticos de 0,99.

3.3 ARTIFICIAL NEURAL NETWORKS FOR SHORT-TERM FORECASTING OF CASES, DEATHS, AND HOSPITAL BEDS OCCUPANCY IN THE COVID-19 PANDEMIC AT THE BRAZILIAN AMAZON

Ainda no contexto da pandemia de COVID-19, (BRAGA *et al.*, 2021) utilizaram-se de um modelo de redes neurais artificiais para predizer tanto o número de casos de COVID-19 quanto a quantidade de mortes e de leitos ocupados. Tendo o estado brasileiro do Pará como objeto de estudo, os mesmos utilizaram dados coletados entre março e junho de 2020. Este artigo se correlaciona fortemente com o presente trabalho, tanto por ter sido desenvolvido com enfoque em uma região brasileira, quanto por utilizar uma técnica relativamente simples e ainda atingir resultados satisfatórios. Entretanto, este trabalho se propõe a utilizar técnicas mais complexas, visando atingir resultados melhores, além disso, (BRAGA *et al.*, 2021) estudou a predição de leitos durante a pandemia, que não será contemplada nesse trabalho, que foca em períodos “normais”.

Foram treinadas diferentes redes neurais para cada objetivo específico, entre eles, a ocupação de leitos hospitalares e leitos de UTI. Levando em consideração seis diferentes cenários. As redes neurais treinadas eram do tipo perceptron multicamadas, adotando uma arquitetura com *feedforward* e uma camada oculta, utilizando-se do algoritmo iterativo Broyden-Fletcher-Goldfarb-Shanno (BROYDEN, 1970) para minimização de erros.

Para a previsão de ocupação de leitos hospitalares, o autores modelaram diversas redes, que possuíam dois neurônios de entrada, de três a nove neurônios na camada intermediária, e dois neurônios de saída. Para a previsão de sete dias, os autores alcançaram uma raiz quadrada do erro médio (Raiz quadrada do erro médio (RMSE)) entre 3,62 e 13,85, dependendo do cenário de treino.

A tabela 3 ilustra um resumo das características principais dos trabalhos correlatos que foram mostrados neste capítulo.

Tabela 3 – Comparação dos Trabalhos Correlatos

Título	**Características**
Predicting intensive care unit bed occupancy for integrated operating room scheduling via neural networks	Utilizam informações extremamente detalhadas sobre os pacientes e sua estadia para prever a ocupação de leitos, utilizando RNAs.
COVID-19 ICU demand forecasting: A two-stage Prophet-LSTM approach	Utilizam dois tipos de ML, Prophet e LSTM, para previsão da ocupação de leitos durante a pandemia de COVID-19.
Artificial neural networks for short-term forecasting of cases, deaths, and hospital beds occupancy in the COVID-19 pandemic at the Brazilian Amazon	Utilizam RNAs para prever casos de COVID-19, mortes e ocupação de leitos no estado do Pará.

Fonte: Elaboração própria

4 DESENVOLVIMENTO

A proposta desse trabalho é fazer um estudo preliminar da aplicação de redes neurais recorrentes do tipo *Long Short-Term Memory* ao problema da ocupação de leitos de UTI. Inicialmente, será feita uma análise dos dados disponibilizados pelo SIH/SUS, visando encontrar um subconjunto que possa garantir as informações necessárias para predição. Serão utilizados apenas dados referentes ao estado de Santa Catarina entre 2010 e 2018.

Após feita a coleta de dados, estes serão pré processados, normalizados e aplicados no treinamento de redes LSTM, que foram descritas no capítulo de fundamentação teórica deste trabalho (capítulo 2). Os resultados dados pelos modelos treinados serão analisados e, a partir desta análise, serão propostas maneiras possíveis de refinar os modelos.

4.1 DADOS UTILIZADOS

Como disposto na seção 2.1, o DATASUS disponibiliza livremente os dados reduzidos das AIH. Para o escopo desse trabalho, apenas algumas informações foram consideradas importantes, são elas:

1. Data de hospitalização
2. Data de alta
3. Número de dias na UTI
4. Tipo de UTI

Como as datas exatas de entrada e saída de UTI dos pacientes não estão disponíveis, não é possível determinar com uma precisão diária quantos pacientes estavam ocupando leitos de UTI, portanto, foi necessário utilizar alguma estratégia para arbitrariamente decidir quando considerar que alguém estava ocupando um leito. A estratégia decidida foi considerar que o paciente é admitido na UTI no mesmo dia que é hospitalizado, e que permanece nela pelo número de dias associados a estadia na UTI, quando então recebe alta ou é transferido para um leito comum, conforme os dados de internação. Essa perda de precisão fez com que a estratégia de agregação de dados diária fosse deixada de lado em favor de uma estratégia de agregação de dados com granularidade de semanas. A estratégia semanal foi escolhida pois a média de tempo de internação dos pacientes em leitos de UTI é de aproximadamente 6,7 dias, enquanto que a mediana é de 4 dias, ou seja, grande parte das pessoas ocupa um leito de UTI por quase uma semana. Entretanto, uma granularidade maior dos dados ainda seria benéfica.

Os dados do SIH disponibilizados pelo DATASUS estão no formato .dbc, que é um formato proprietário do Departamento de Informática do SUS. Logo, nenhuma ferramenta

de visualização ou processamento de dados do mercado consegue interpretar esses dados puros. Portanto, é necessário que seja feita uma conversão para algum outro formato legível pelas ferramentas disponíveis no mercado.

A linguagem de programação R foi escolhida como ferramenta para conversão desses dados, pois a mesma conta com um pacote de leitura de arquivos no formato .dbc, o read.dbc, que atualmente está disponível no CRAN[1]. O pacote converte os arquivos .dbc para um *dataframe*, estrutura de dados padrão da linguagem, que pode ser salvo como diversos formatos, como por exemplo CSV. O código utilizado para coleta e conversão dos dados foi adaptado a partir da documentação sobre o Sistema de Informações Hospitalares do SUS (CIÊNCIA DE DADOS APLICADA À SAÚDE, 2019).

Uma vez convertidos os dados para o formato CSV, foi utilizada a linguagem Python, junto com a biblioteca Pandas, para fazer o pré-processamento dos dados para posterior análise. As colunas que não possuíam dados de interesse foram removidas, então os dados foram agrupados de maneira semanal para análise e treinamento das redes neurais.

O pré-processamento dos dados para análise preliminar consistiu dos seguintes passos:

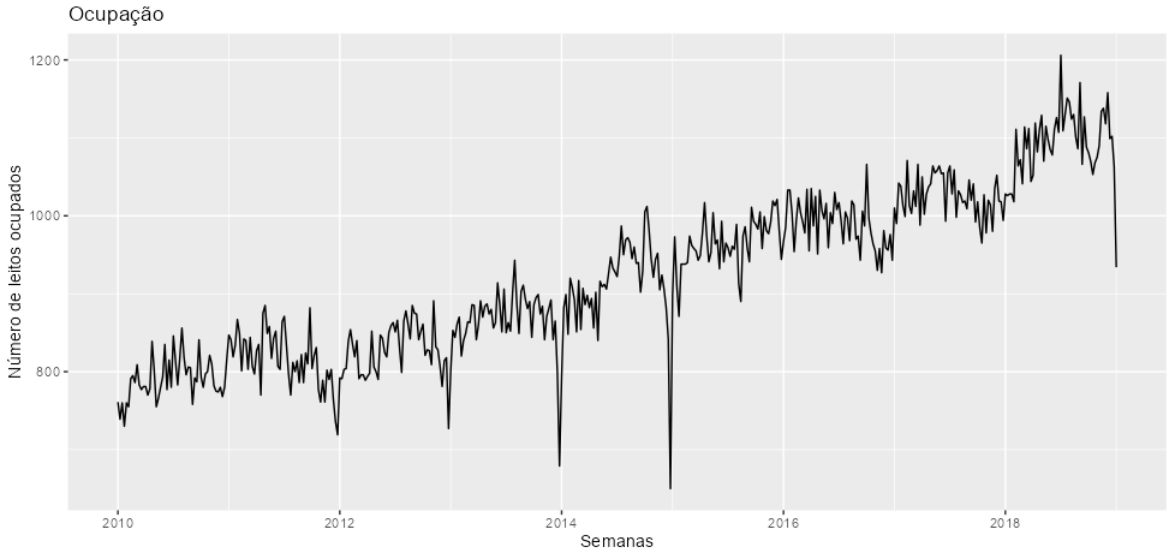
1. Selecionar apenas as colunas de interesse;
2. Transformar as datas do tipo string para o formato datetime do Pandas;
3. Criar uma nova coluna para armazenar a data de alta da UTI;
4. Calcular a data de saída da UTI, utilizando a data da hospitalização como data de entrada na UTI e somando o número de dias de estadia na UTI (diminuindo de um, já que o dia que o paciente foi hospitalizado também entra na conta);
5. Agrupar os registros por semanas, sendo que é considerado que se a pessoa estava qualquer quantidade de tempo na UTI em uma dada semana, ela estava ocupando um leito de UTI naquela semana;
6. Contar a quantidade de registros em cada semana e criar uma nova tabela apenas com o número da semana e a quantidade de leitos ocupados.

4.1.1 Análise preliminar dos dados

A figura 7 mostra a ocupação de leitos de UTI semanal entre a primeira semana do ano de 2010 e a última semana do ano de 2018, já é possível verificar visualmente que existe uma certa tendência de aumento do número da ocupação dos leitos de UTI, assim como uma certa sazonalidade, com um aumento no número de casos no inverno, e uma diminuição no verão.

[1] Rede de distribuição de pacotes do R

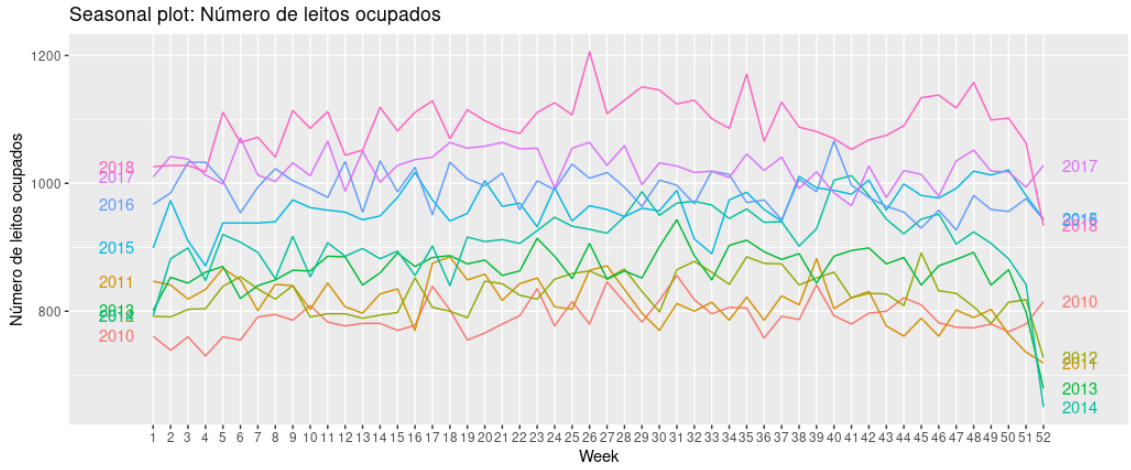
Figura 7 – Ocupação de leitos de UTI semanal entre 2010 e 2018



Fonte: Elaboração própria a partir dos dados do SIH/SUS

O gráfico mostrado na figura 8 apresenta uma comparação da ocupação semanal de leitos de UTI entre os anos analisados. A figura demonstra uma similaridade entre os comportamentos dos anos, além disso, ela reforça a noção de que cada ano passado existe um aumento na ocupação de leitos, explicitado pelo dos valores a cada ano. Para tentar explicar esse comportamento, foram buscados dados do Instituto Brasileiro de Geografia e Estatística sobre a população do estado de Santa Catarina de 2010, assim como sua projeção para os anos seguintes (IBGE, D. d. P., 2023).

Figura 8 – Comparação da ocupação de leitos de UTI semanal entre 2010 e 2018

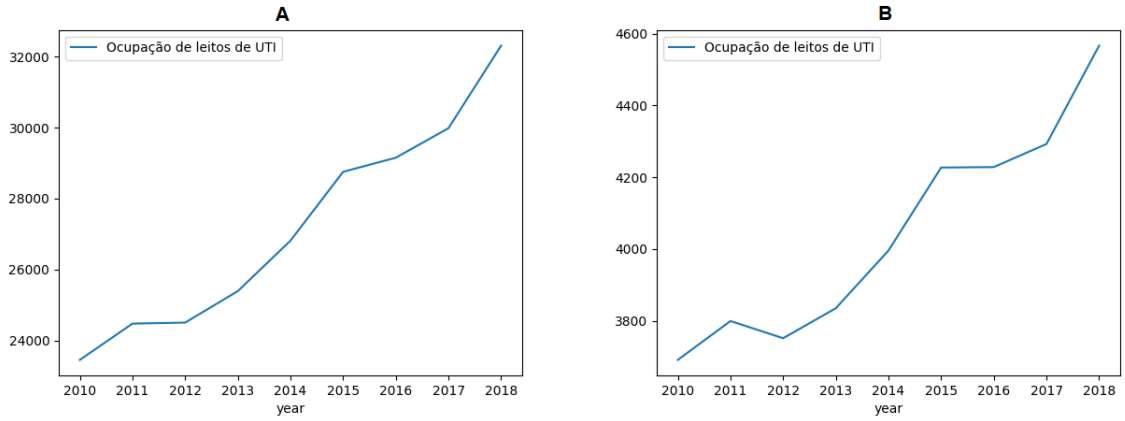


Fonte: Elaboração própria a partir dos dados do SIH/SUS

Usando esses dados, dois gráficos foram produzidos: o gráfico A, com os números

absolutos de ocupação de leitos para cada ano; o gráfico B, com o número de ocupação de leitos em relação à estimativa populacional daquele ano. A figura 9 apresenta a comparação dos gráficos, pode-se observar que apesar de existir um pequeno achatamento da curva, o crescimento da ocupação de leitos ainda é extremamente evidente, e portanto, o crescimento populacional não explica completamente o aumento paulatino da ocupação de leitos.

Figura 9 – Comparação entre números de ocupação de leitos de UTI anuais absolutos, e relativos ao crescimento populacional

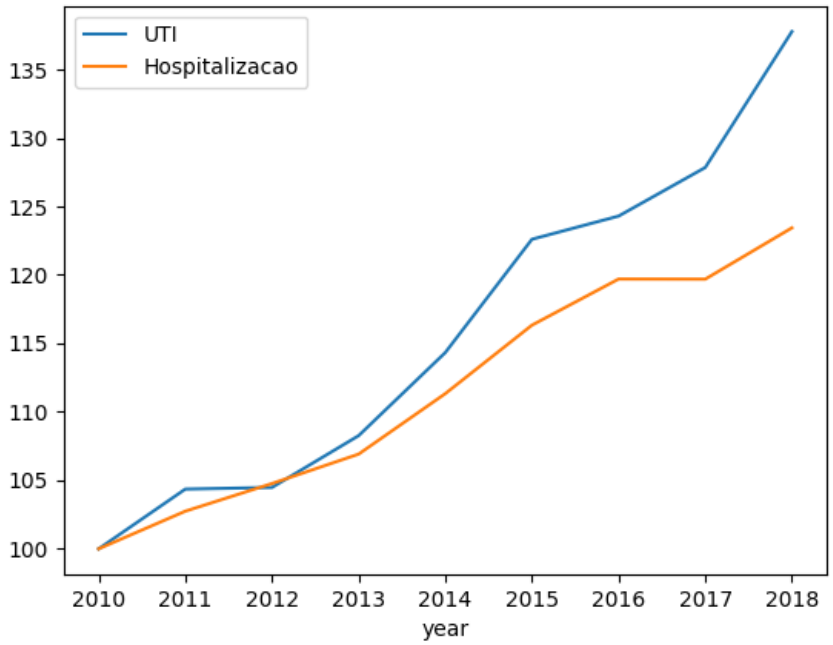


Fonte: Elaboração própria a partir dos dados do SIH/SUS

Também foi verificado se este aumento da demanda por leitos de UTI acompanha um aumento da demanda por hospitalizações em geral, entretanto, como pode-se observar na figura 10, enquanto que a demanda por hospitalizações cresceu em torno de 22%, a demanda por leitos de UTI cresceu 35%. É provável que outros fatores externos causem essa tendência de crescimento da demanda por leitos de UTI, como aumento da oferta de leitos ou envelhecimento da população, entretanto, uma análise mais aprofundada sobre essa causa foge do escopo deste trabalho.

Finalmente, foi feita uma decomposição da série temporal, utilizando o método de decomposição de Loess, baseada na função de suavização local ponderada por polinômios (CLEVELAND *et al.*, 1990), a mesma encontra-se na figura 11. A decomposição da série temporal está dividida em três painéis (tendência, sazonalidade e resíduo), cada um com uma componente da série. Esses componentes podem ser somados para reconstruir os dados mostrados no painel superior (data). Observa-se que o componente sazonal não apresenta mudança significativa ao longo do tempo, de modo que qualquer ano consecutivo apresenta padrão semelhante. O componente de resíduo mostrado no painel inferior é o que resta quando os componentes sazonal e de tendência-ciclo são subtraídos dos dados. As barras cinzas à esquerda de cada painel mostram as escalas relativas dos componentes. Cada barra cinza representa o mesmo comprimento, mas devido às diferentes escalas dos

Figura 10 – Comparação do crescimento da demanda por leitos de UTI e da demanda de hospitalizações

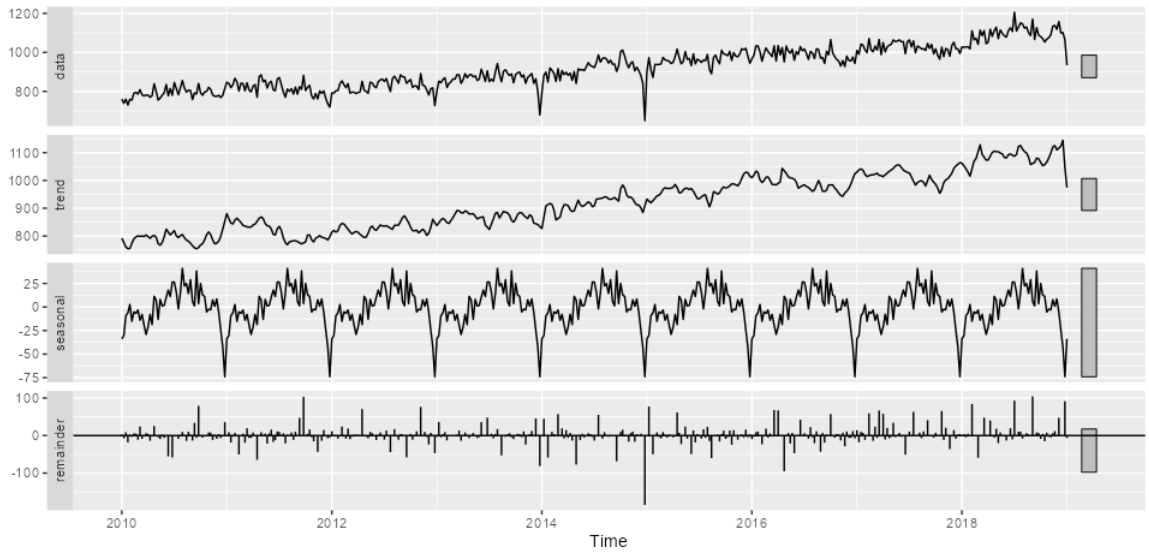


Fonte: Elaboração própria a partir dos dados do SIH/SUS

gráficos, as barras variam em tamanho. Analisando a figura, é possível confirmar que existe uma tendência clara de crescimento da série, assim como uma sazonalidade evidente, que tende a levar à um aumento dos valores da série no meio do ano.

A análise dos dados serviu para confirmar que a série tratada tem um comportamento dentro do esperado de uma série temporal comum, isto é, possui uma sazonalidade clara, além de uma tendencia de crescimento quase constante, com resíduos relativamente pequenos, ou seja, todas essas características mostram que a série não segue um padrão aleatório, mas possuí um mecanismo por trás que pode eventualmente ser aproximado por alguma técnica, como redes neurais, de maneira que possamos atingir predições com alta acurácia. Ademais, é interessante que a análise seja feita antes do início do treinamento das redes, pois desta forma, podemos criar uma intuição de quais seriam as melhores maneiras de treinar a rede. Por exemplo, como a autocorrelação da série é relativamente alta, podemos tentar utilizar desde o início mais de uma semana como entrada da rede.

Figura 11 – Decomposição da série temporal



Fonte: Elaboração própria a partir dos dados do SIH/SUS

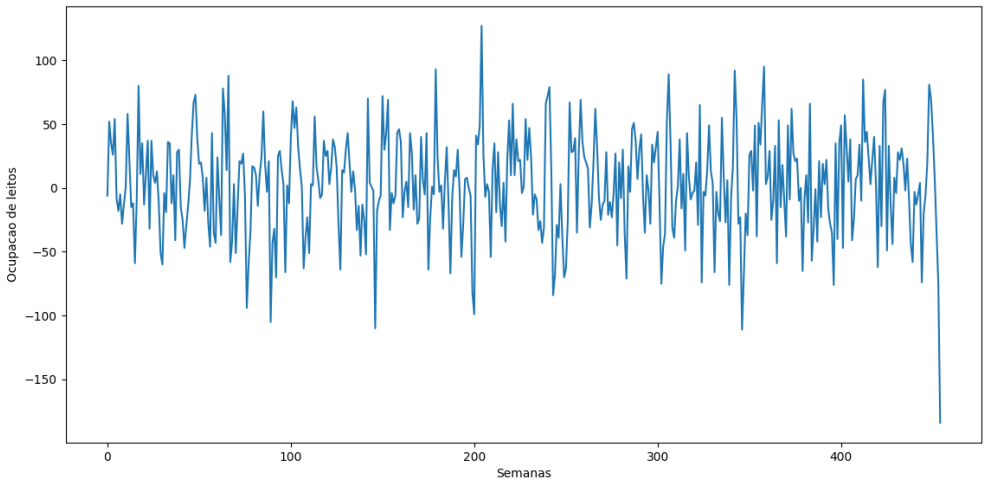
4.2 TREINAMENTO DAS REDES LSTM

Antes de iniciar o treinamento de redes neurais é necessário fazer um tratamento dos dados, afim de extrair a melhor performance das mesmas. O primeiro passo do tratamento foi a remoção dos *outliers* estatísticos, para isso foi feito o cálculo dos quantis de 1% e 99% dos dados, valores fora dessa faixa foram descartados.

O próximo passo foi transformar a série temporal em estacionária, removendo a tendência da mesma, um procedimento comum para isso é tomar a diferença da série(COGHLAN, 2015), isto é, cada valor é subtraído do anterior, de forma que a série temporal resultante representa as diferenças entre os valores originais. O resultado desse processo pode ser observado na figura 12. Ademais, os dados foram normalizados entre -1 e 1, outro procedimento comum para o treinamento de redes neurais, a normalização foi feita com auxílio da biblioteca *open source* scikit-learn do Python.

Para que a rede treinada utilize mais do que apenas uma semana como entrada para predizer a próxima, ainda é necessário fazer mais uma última transformação dos dados de entrada. Precisamos modificar a dimensão do vetor de entrada da rede, para que o mesmo possua a quantidade correta de semanas por tupla. Por exemplo, para que seja possível utilizar as últimas três semanas para predizer a próxima, é necessário que cada tupla do vetor de entrada possua três valores, um com o valor da ocupação no tempo $T-2$ outro no tempo $T-1$ e finalmente um no tempo T. Na biblioteca Pandas, primeiramente a quantidade de colunas necessária é criada no *dataframe* de entrada, então cada coluna é populada com os valores necessários.

Figura 12 – Série estacionária



Fonte: Elaboração própria a partir dos dados do SIH/SUS

A função de ativação utilizada foi a tangente hiperbólica, a função recorrente[2] utilizada foi a sigmoide logística, ambas as funções são comumente usadas em redes LSTM, como apresentado na seção 2.3.2. O otimizador escolhido foi o Adam, pois o mesmo é robusto e muito utilizado em uma gama de problemas de otimização na área de aprendizado de máquina(KINGMA; BA, 2014). A função de perda utilizada foi o erro médio quadrático. Todas as redes foram treinadas com 300 épocas, utilizando *early stopping* com paciência de 6, de forma que nenhuma das redes chegou a realmente treinar pelo número máximo de épocas. O conjunto de treinamento sempre representava 75% dos dados, enquanto que o conjunto de testes era composto pelo restante. Como as redes LSTM possuem um estado interno, entre o fim do treinamento e o início dos testes, as redes foram alimentadas com os dados de treinamento uma última vez, para que a rede possuísse um estado interno construído antes de começar os testes.

A medida de acurácia escolhida utilizada para julgar a qualidade das redes treinadas foi a raiz do erro quadrático médio, ou RMSE. Como o efeito de cada erro no RMSE é proporcional ao erro quadrático, erros maiores tem um efeito desproporcionalmente grande no RMSE, fazendo com que o RMSE seja mais sensível à *outliers*, para o objeto de estudo deste trabalho, é extremamente importante que mesmo os *outliers* sejam previstos com uma acurácia decente, portanto a escolha da medida de acurácia se justifica. Além disso, o RMSE sempre está na mesma escala dos dados, portanto serve apenas para comparações dentro do mesmo conjunto de dados. Historicamente o RMSE sempre foi uma medida popular, principalmente por causa da sua relevância teórica na modelagem estatística

[2] A função de ativação dos portões de entrada, saída e esquecimento das células.

(HYNDMAN; KOEHLER, 2006). A métrica secundária de comparação utilizada foi o MAPE, ou erro percentual absoluto médio, utilizado na comparação entre os modelos desenvolvidos.

Foram criados diversos modelos com intenção de buscar qual tipo de modelagem levaria à uma acurácia melhor das redes LSTM. Além disso, para cada modelo, os hiperparâmetros foram ajustados manualmente na tentativa de melhorar o desempenho das redes. O primeiro modelo foi o modelo "generalista", que utiliza todos os dados disponíveis para treinamento e validação. Os demais trabalharam apenas com uma parte dos dados, com a intenção explorar possíveis similaridades, foram criados modelos treinados separados por períodos de governos de presidentes da república, além de modelos treinados apenas por certos períodos do ano, como outono e inverno.

4.2.1 Modelo generalista

Como apresentado anteriormente, o primeiro modelo treinado utiliza indiscriminadamente todos os dados disponíveis, ou seja, os dados de 2010 até 2016 foram utilizados para treinamento da rede, enquanto que os dados de 2017 até 2018 foram utilizados para a validação da mesma. A tabela 4 apresenta os resultados das redes que tiveram maior desempenho, na primeira coluna é mostrada a quantidade de semanas antecedentes utilizadas para prever a semana seguinte, a segunda coluna apresenta a arquitetura da rede[3] e a terceira coluna o RMSE.

Tabela 4 – Modelo generalista

Semanas consideradas	**Arquitetura da rede**	**RMSE alcançado**
4 semanas anteriores	4 neurônios	46,928
3 semanas anteriores	4 neurônios	46,977
3 semanas anteriores	3 neurônios	48,794
2 semanas anteriores	4 neurônios	51,154
4 semanas anteriores	**4/2 neurônios**	**43,408**
3 semanas anteriores	4/2 neurônios	48,209
2 semanas anteriores	4/2 neurônios	52,529

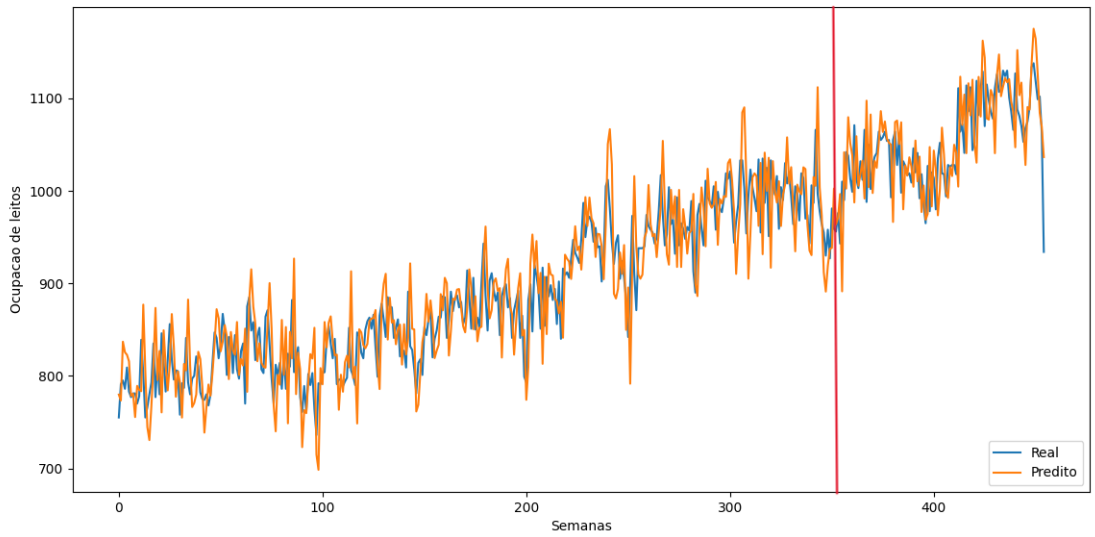
Fonte: Elaboração própria

É possível ver que os RMSEs alcançados são relativamente próximos, entretanto, redes que consideram uma quantidade de semanas maior que dois performam relativamente melhor, com a melhor rede considerando quatro semanas anteriores e possuindo duas camadas com quatro e dois neurônios, respectivamente. Durante o treinamento, redes que consideravam uma quantidade maior que quatro semanas foram treinadas, porém após a quarta semana, o desempenho das redes começa a cair significativamente, o que indica

[3] Quando a rede possui mais de uma *hidden layer* as camadas são representadas como X/Y, onde X e Y são a quantidade de neurônios nas camadas um e dois, respectivamente.

que a autocorrelação da série temporal tende a cair consideravelmente após um atraso de quatro semanas.

Figura 13 – Gráfico de predições do modelo generalista



Fonte: Elaboração própria

A figura 13 mostra um gráfico comparando os valores da série temporal completa, e os valores preditos pela rede que obteve o melhor desempenho para o modelo, os valores anteriores à linha vermelha sendo valores do conjunto de treinamento, e os valores posteriores sendo do conjunto de testes.

4.2.2 Modelos separados por governos federais

Os próximos três modelos criados segregam os dados nos respectivos períodos de atuação dos presidentes da república, a ideia por trás dessa separação é que cada governo tende a ter certas políticas públicas sobre a saúde, o que poderia portanto impactar o comportamento da série temporal, de maneira que, redes treinadas especificamente para um certo governo poderiam ter um desempenho mais satisfatório.

O primeiro modelo treinado considera os primeiros quatro anos do período dos dados coletados, refletindo o período do primeiro governo da então presidente da república Dilma Rousseff, o segundo modelo reflete o período do segundo governo da ex-presidente, até seu eventual término via impeachment, o último modelo dentre os três reflete o período de governo do então presidente Michel Temer.

A tabela 5 apresenta os melhores resultados obtidos no treinamento das redes que seguem o modelo do primeiro governo, é possível perceber que houve uma melhora na acurácia das redes treinadas em relação ao modelo genérico, possivelmente pois a conjectura levantada anteriormente realmente tem uma certa validade, ou seja, é provável

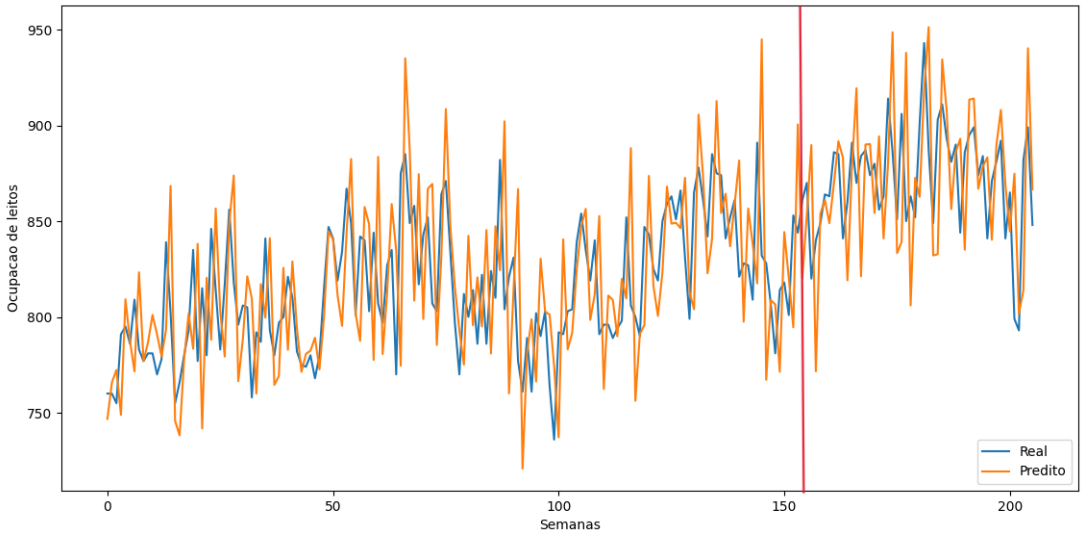
Tabela 5 – Modelo do primeiro governo

Semanas consideradas	**Arquitetura da rede**	**RMSE alcançado**
4 semanas anteriores	4 neurônios	47,107
3 semanas anteriores	4 neurônios	42,311
3 semanas anteriores	3 neurônios	53,142
2 semanas anteriores	4 neurônios	44,498
4 semanas anteriores	4/2 neurônios	45,891
3 semanas anteriores	4/2 neurônios	43,072
2 semanas anteriores	**4/2 neurônios**	**41,570**

Fonte: Elaboração própria

que redes treinadas considerando as políticas específicas de certo governo tenham uma acurácia melhor na previsão da ocupação de leitos.

Figura 14 – Gráfico de predições do primeiro governo



Fonte: Elaboração própria

A figura 14 mostra um gráfico comparando os valores da série temporal durante período do primeiro governo, e os valores preditos pela rede que obteve o melhor desempenho para o modelo, os valores anteriores à linha vermelha sendo valores do conjunto de treinamento, e os valores posteriores sendo do conjunto de testes.

Os resultados obtidos para as redes treinadas no segundo governo do período, apresentados na tabela 6, foram os piores dentre todos os modelos, uma possível explicação é que a instabilidade política do período tenha afetado o comportamento da série, fazendo com que a mesma seja de mais difícil interpretação e previsão pelas redes LSTM.

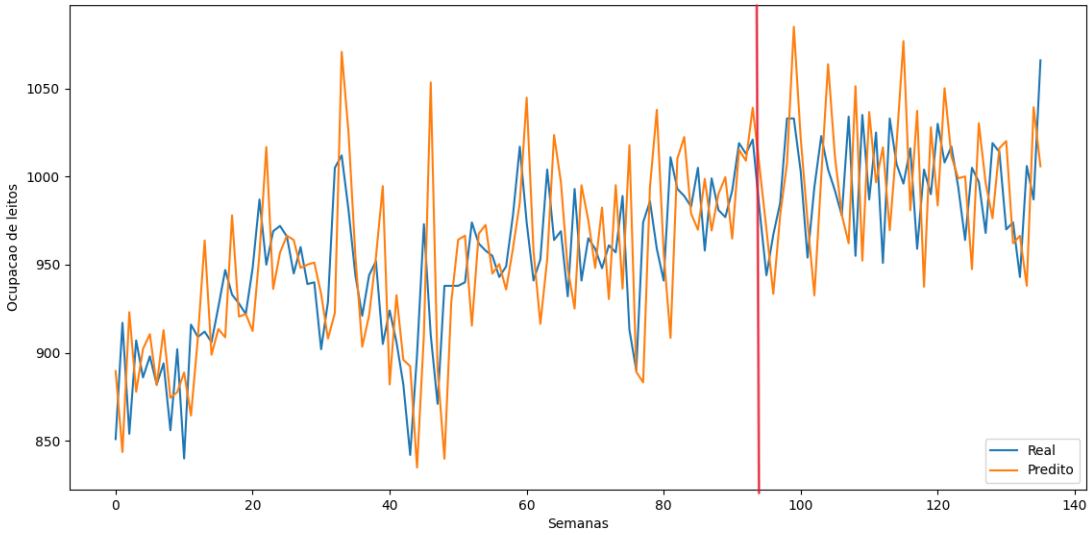
A figura 15 mostra um gráfico comparando os valores da série temporal durante o período do segundo governo, e os valores preditos pela rede que obteve o melhor desempe-

Tabela 6 – Modelo do segundo governo

Semanas consideradas	**Arquitetura da rede**	**RMSE alcançado**
4 semanas anteriores	4 neurônios	62,599
3 semanas anteriores	**4 neurônios**	**50,663**
3 semanas anteriores	3 neurônios	52,195
2 semanas anteriores	4 neurônios	64,488
4 semanas anteriores	4/2 neurônios	61,967
3 semanas anteriores	4/2 neurônios	51,170
2 semanas anteriores	4/2 neurônios	61,633

Fonte: Elaboração própria

Figura 15 – Gráfico de predições do segundo governo



Fonte: Elaboração própria

nho para o modelo, os valores anteriores à linha vermelha sendo valores do conjunto de treinamento, e os valores posteriores sendo do conjunto de testes.

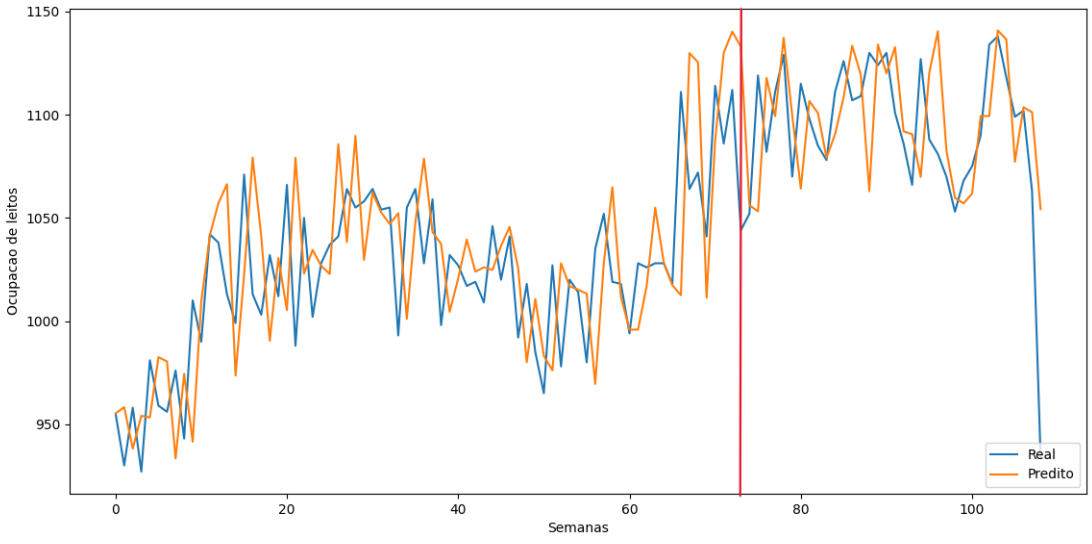
Tabela 7 – Modelo do terceiro governo

Semanas consideradas	**Arquitetura da rede**	**RMSE alcançado**
4 semanas anteriores	4 neurônios	41,745
3 semanas anteriores	**4 neurônios**	**37,038**
3 semanas anteriores	3 neurônios	40,456
2 semanas anteriores	4 neurônios	40,967
4 semanas anteriores	4/2 neurônios	42,603
3 semanas anteriores	4/2 neurônios	39,411
2 semanas anteriores	4/2 neurônios	40,452

Fonte: Elaboração própria

A tabela 7 apresenta os melhores resultados obtidos nos treinamentos das redes seguindo o modelo do terceiro governo, dentre todos os modelos feitos, as redes deste modelo foram as que previram melhor a ocupação de leitos de UTI, segundo a métrica adotada. Isto pode ser dado devido a uma combinação dos fatores anteriormente mencionados, ou seja, as políticas de governo na saúde estabilizaram a série, além disso, a instabilidade política foi diminuída.

Figura 16 – Gráfico de predições do terceiro governo



Fonte: Elaboração própria

A figura 16 mostra um gráfico comparando os valores da série temporal durante o período do terceiro governo, e os valores preditos pela rede que obteve o melhor desempenho para o modelo, os valores anteriores à linha vermelha sendo valores do conjunto de treinamento, e os valores posteriores sendo do conjunto de testes.

4.2.3 Modelos separados por estações do ano

Como foi apresentado na seção 4.1.1, a série temporal que é objeto de estudo deste trabalho possui uma sazonalidade significativa, levando isso em conta, os últimos dois modelos criados focam em separar o ano em duas partes, outono/inverno e primavera/verão, com a intenção de treinar redes que são especializadas em predizer o comportamento da ocupação de leitos durante certas estações.

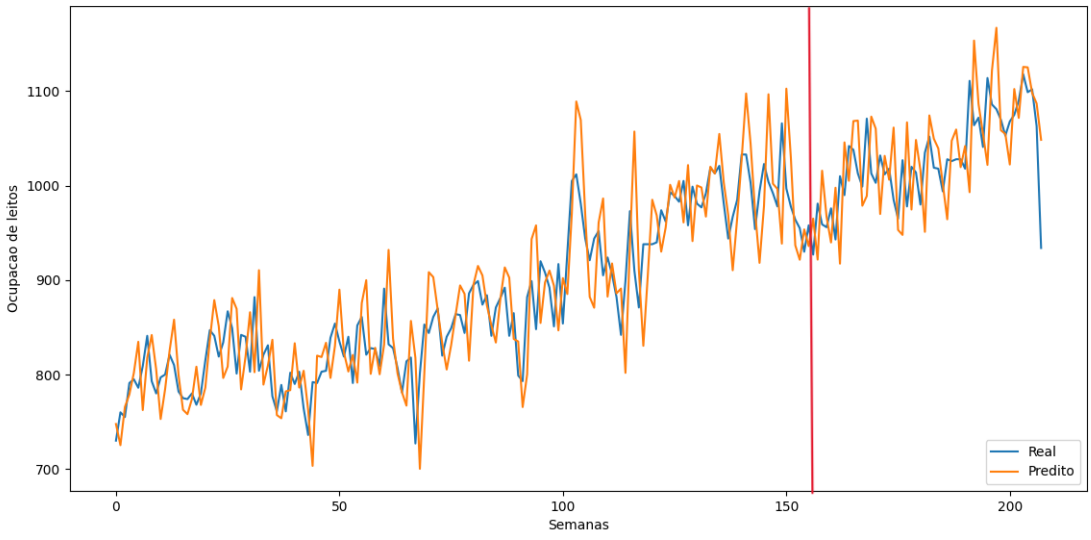
Tabela 8 – Modelo primavera/verão

Semanas consideradas	**Arquitetura da rede**	**RMSE alcançado**
4 semanas anteriores	4 neurônios	53,827
3 semanas anteriores	**4 neurônios**	**50,606**
3 semanas anteriores	3 neurônios	51,004
2 semanas anteriores	4 neurônios	52,202
4 semanas anteriores	4/2 neurônios	51,302
3 semanas anteriores	4/2 neurônios	52,189
2 semanas anteriores	4/2 neurônios	54,166

Fonte: Elaboração própria

Analisando a tabela 8, percebe-se que os resultados são relativamente ruins para todas as redes treinadas, parte do motivo pode ser o fato de que em vários dos anos estudados, existe uma oscilação grande no número registrado de pacientes ocupando leitos na última semana do ano, que pode afetar negativamente o treino da rede. Outra possibilidade é que a sazonalidade no período da primavera e verão é pouco pronunciada, e portanto de mais difícil interpretação pelas redes neurais.

Figura 17 – Gráfico de predições do modelo primavera/verão



Fonte: Elaboração própria

A figura 13 mostra um gráfico comparando os valores da série temporal durante o período de primavera/verão, e os valores preditos pela rede que obteve o melhor desempenho para o modelo, os valores anteriores à linha vermelha sendo valores do conjunto de treinamento, e os valores posteriores sendo do conjunto de testes.

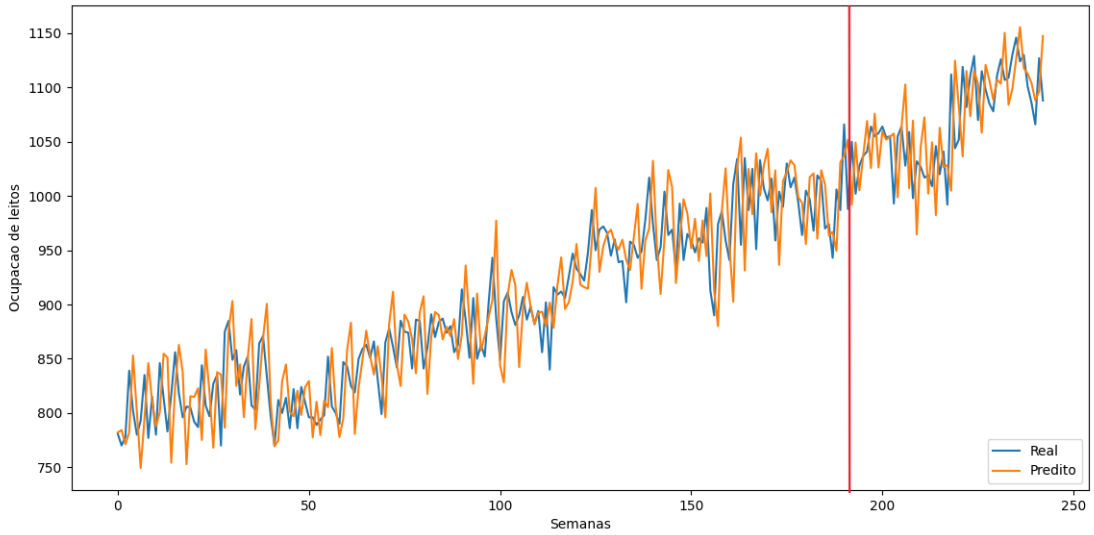
Tabela 9 – Modelo outono/inverno

Semanas consideradas	**Arquitetura da rede**	**RMSE alcançado**
4 semanas anteriores	4 neurônios	46,246
3 semanas anteriores	**4 neurônios**	**42,377**
3 semanas anteriores	3 neurônios	52,799
2 semanas anteriores	4 neurônios	54,168
4 semanas anteriores	4/2 neurônios	52,257
3 semanas anteriores	4/2 neurônios	45,362
2 semanas anteriores	4/2 neurônios	52,884

Fonte: Elaboração própria

Finalmente, analisando a tabela 9, percebe-se que os resultados são relativamente parecidos com os do modelo genérico, levando em conta isso e os resultados do modelo da primavera e verão, pode-se afirmar que modelos que levam em conta apenas estações específicas do ano não parecem ter um desempenho significativamente melhor do que modelos que ignoram isso.

Figura 18 – Gráfico de predições do modelo outono/inverno



Fonte: Elaboração própria

A figura 18 mostra um gráfico comparando os valores da série temporal durante o período de outono/inverno, e os valores preditos pela rede que obteve o melhor desempenho para o modelo, os valores anteriores à linha vermelha sendo valores do conjunto de treinamento, e os valores posteriores sendo do conjunto de testes.

4.2.4 Comparação entre as redes desenvolvidas

A tabela 10 demonstra uma comparação entre as melhores redes desenvolvidas, além de trazer o RMSE das redes, o Erro Percentual Absoluto Médio (MAPE) também é mostrado. É possível ver que a rede do terceiro governo se saiu muito melhor, mesmo na métrica secundária adotada, o MAPE, alcançando um erro percentual de 2,434%.

Tabela 10 – Melhores redes de cada agrupamento

Agrupamento	**Semanas consideradas**	**Arquitetura da rede**	**RMSE**	**MAPE**
Generalista	4 semanas	4/2 neurônios	43,408	3,392
Primeiro Governo	2 semanas	4/2 neurônios	41,570	3,836
Segundo Governo	3 semanas	4 neurônios	50,663	4,090
Terceiro Governo	**3 semanas**	**4 neurônios**	**37,038**	**2,434**
Primavera/Verão	3 semanas	4 neurônios	50,606	4,170
Outono/Inverno	3 semanas	4 neurônios	42,377	3,432

Fonte: Elaboração própria

5 CONCLUSÃO

Este trabalho apresentou um estudo preliminar sobre o comportamento da ocupação de leitos hospitalares de UTI no estado de Santa Catarina, explicando o comportamento da série temporal de ocupações de leitos, como visto na seção 4.1.1, a ocupação de leitos segue uma tendência clara de crescimento, que não é explicada apenas pelo crescimento populacional, além disso, existe um componente sazonal importante que influencia o comportamento da série. A causa da tendência de crescimento continua desconhecida, e portanto, estudos futuros podem buscar descobrir qual a fonte dessa tendência.

Além disso, este documento apresentou o uso de rede do tipo Long Short Term Memory para o problema de predição de ocupação de leitos, diversas configurações foram feitas, entre elas, foram utilizados de 3 a 6 neurônios, até duas camadas ocultas e considerando de 2 a 4 semanas anteriores para previsão, além disso, diversas variações no conjunto de dados utilizados foram feitas, primeiramente criando um modelo generalista, com todos os dados, e posteriormente, separando os dados por governos e sazonalidade. Foram apresentadas as motivações para cada modelo de rede proposta, assim como comparações entre os modelos, com o objetivo de buscar quais modelos foram os mais bem sucedidos.

Desta forma, com base nos dados disponíveis, é possível chegar a conclusão de que enquanto um modelo genérico que ignora as nuances dos dados pode ter uma boa acurácia, modelos construídos para situações mais específicas, como por exemplo modelos que consideram apenas um governo, tendem a apresentar desempenhos melhores. O modelo mais bem sucedido foi o modelo que trabalhou sobre os dados do terceiro governo do período, possuindo apenas uma camada oculta com 4 unidades LSTM e considerando 3 semanas anteriores para a previsão, alcançando um RMSE de 37,038 e um MAPE de 2,434. É importante levar em conta, entretanto, que a falta de dados mais precisos dificultou alguns objetivos do trabalho. A falta de uma granularidade diária nos dados tornou a rede menos precisa, assim como a falta de algumas informações como o número total de leitos, que poderia servir de limite superior para as predições. Trabalhos futuros devem buscar fontes de dados mais ricas, para que estudos mais aprofundados, que levem em conta mais características do problema, possam ser feitos.

Trabalhos futuros também podem focar em fazer uso das redes desenvolvidas no presente trabalho, para criação de ferramentas *user friendly*, como uma interface *web*, por exemplo, que mostre os dados coletados, assim como os resultados da predições, e permita a inserção de novos dados. Além disso, outros trabalhos futuros podem focar em fazer comparações entre as redes desenvolvidas neste trabalho e outras técnicas de predição disponíveis na literatura, utilizadas sobre o mesmo conjunto de dados. Uma vez que este tipo de predição lida com a vida de pacientes, é de extrema importância que a melhor acurácia possível para as predições seja buscada, independente da técnica utilizada.

REFERÊNCIAS

ABIODUN, Oludare Isaac *et al.* State-of-the-art in artificial neural network applications: A survey. **Heliyon**, Elsevier, v. 4, n. 11, e00938, 2018.

ACADEMY, Data Science. **Deep Learning Book**. Disponível em: https://www.deeplearningbook.com.br/arquitetura-de-redes-neurais-long-short-term-memory/. Acesso em: 9 jun. 2023.

ALBAWI, Saad; MOHAMMED, Tareq Abed; AL-ZAWI, Saad. Understanding of a convolutional neural network. *In:* 2017 International Conference on Engineering and Technology (ICET). [*S.l.*: *s.n.*], 2017. P. 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186.

ALPAYDIN, Ethem. **Introduction to machine learning**. [*S.l.*]: MIT press, 2020.

AMARI, Shun-ichi. Backpropagation and stochastic gradient descent method. **Neurocomputing**, Elsevier, v. 5, n. 4-5, p. 185–196, 1993.

BEENHAKKER, Henri L. Multiple correlation—A technique for prediction of future hospital bed needs. **Operations Research**, INFORMS, v. 11, n. 5, p. 824–839, 1963.

BISHOP, Chris M. Neural networks and their applications. **Review of scientific instruments**, American Institute of Physics, v. 65, n. 6, p. 1803–1832, 1994.

BORGES, Dalton; NASCIMENTO, Mariá CV. COVID-19 ICU demand forecasting: A two-stage Prophet-LSTM approach. **Applied Soft Computing**, Elsevier, v. 125, p. 109181, 2022.

BOVIM, Thomas Reiten *et al.* Stochastic master surgery scheduling. **European Journal of Operational Research**, v. 285, n. 2, p. 695–711, 2020. ISSN 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2020.02.001. Disponível em: https://www.sciencedirect.com/science/article/pii/S0377221720301041.

BRAGA, Marcus de Barros *et al.* Artificial neural networks for short-term forecasting of cases, deaths, and hospital beds occupancy in the COVID-19 pandemic at the Brazilian Amazon. **PLoS One**, Public Library of Science San Francisco, CA USA, v. 16, n. 3, e0248161, 2021.

BROCKWELL, Peter J; DAVIS, Richard A. **Introduction to time series and forecasting**. [*S.l.*]: Springer, 2002.

BROYDEN, C. G. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. **IMA Journal of Applied Mathematics**, v. 6, n. 1, p. 76–90, mar. 1970. ISSN 0272-4960. DOI: 10.1093/imamat/6.1.76. eprint:

https://academic.oup.com/imamat/article-pdf/6/1/76/2233756/6-1-76.pdf. Disponível em: https://doi.org/10.1093/imamat/6.1.76.

CARVALHO, Déa M. T. Sistema de Informações Hospitalares do SUS – SIH-SUS. **A experiência brasileira em sistemas de informação em saúde**, v. 49, 2009.

CHATFIELD, Christopher. **The analysis of time series: theory and practice**. [*S.l.*]: Springer, 2013.

CIÊNCIA DE DADOS APLICADA À SAÚDE, Plataforma de. **Sistema de Informações Hospitalares do SUS – SIHSUS**. 2019. Disponível em: https://pcdas.icict.fiocruz.br/conjunto-de-dados/sistema-de-informacoes-hospitalares-do-sus-sihsus/documentacao/. Acesso em: 10 fev. 2023.

CLEVELAND, Robert B *et al.* STL: A seasonal-trend decomposition. **J. Off. Stat**, v. 6, n. 1, p. 3–73, 1990.

COGHLAN, Avril. A little book of R for time series. **Published under Creative Commons Attribution**, v. 3, 2015.

CRYER, Jonathan D. **Time series analysis**. [*S.l.*]: Duxbury Press Boston, 1986. v. 286.

EARNEST, Arul *et al.* Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. **BMC Health Services Research**, BioMed Central, v. 5, n. 1, p. 1–8, 2005.

ESLING, Philippe; AGON, Carlos. Time-series data mining. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 45, n. 1, p. 1–34, 2012.

GERS, Felix A; SCHRAUDOLPH, Nicol N; SCHMIDHUBER, Jürgen. Learning precise timing with LSTM recurrent networks. **Journal of machine learning research**, v. 3, Aug, p. 115–143, 2002.

GRAVES, Alex. Long short-term memory. **Supervised sequence labelling with recurrent neural networks**, Springer, p. 37–45, 2012.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

HYNDMAN, Rob J; ATHANASOPOULOS, George. **Forecasting: principles and practice**. [*S.l.*]: OTexts, 2018.

HYNDMAN, Rob J; KOEHLER, Anne B. Another look at measures of forecast accuracy. **International journal of forecasting**, Elsevier, v. 22, n. 4, p. 679–688, 2006.

IBGE. **Sistema de Informações Hospitalares do SUS – SIH/SUS**. Disponível em: `https://ces.ibge.gov.br/base-de-dados/metadados/ministerio-da-saude/sistema-de-informacoes-hospitalares-do-sus-sih-sus.html`. Acesso em: 11 abr. 2023.

IBGE, Diretoria de Pesquisas. **Projeção da população**. Disponível em: `https://cidades.ibge.gov.br/brasil/sc/panorama`. Acesso em: 13 jun. 2023.

KINGMA, Diederik P; BA, Jimmy. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

KOHONEN, Teuvo. **Self-organization and associative memory**. [*S.l.*]: Springer Science & Business Media, 2012. v. 8.

KUTAFINA, Ekaterina *et al.* Recursive neural networks in hospital bed occupancy forecasting. **BMC medical informatics and decision making**, Springer, v. 19, n. 1, p. 1–10, 2019.

LESSA, Fábio José Delgado *et al.* Novas metodologias para vigilância epidemiológica: uso do Sistema de Informações Hospitalares-SIH/SUS. **Informe Epidemiológico do SUS**, Centro Nacional de Epidemiologia/Fundação Nacional de Saúde/Ministério da Saúde, v. 9, p. 3–19, 2000.

LEVCOVITZ, Eduardo; PEREIRA, Telma Ruth C. SIH/SUS (Sistema AIH): uma análise do sistema público de remuneração de internações hospitalares no Brasil-1983-1991. *In:* SIH/SUS (Sistema AIH): uma análise do sistema público de remuneração de internações hospitalares no Brasil-1983-1991. [*S.l.*: *s.n.*], 1993. P. 83–83.

LITTIG, Steven J; ISKEN, Mark W. Short term hospital occupancy prediction. **Health care management science**, Springer, v. 10, n. 1, p. 47–66, 2007.

MAYER, Hermann *et al.* A system for robotic heart surgery that learns to tie knots using recurrent neural networks. **Advanced Robotics**, Taylor & Francis, v. 22, n. 13-14, p. 1521–1537, 2008.

PHUNG; RHEE. A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets. **Applied Sciences**, v. 9, p. 4500, out. 2019. DOI: `10.3390/app9214500`.

PHUNG, Van Hiep; RHEE, Eun Joo. A deep learning approach for classification of cloud image patches on small datasets. **Journal of information and communication convergence engineering**, The Korea Institute of Information e Commucation Engineering, v. 16, n. 3, p. 173–178, 2018.

ROBINSON, AJ; FALLSIDE, Frank. **The utility driven dynamic error propagation network**. [*S.l.*]: University of Cambridge Department of Engineering Cambridge, 1987.

ROISENBERG, Mauro. **Introdução ao Estudo das Redes Neurais Artificiais**. [*S.l.*: *s.n.*]. `http://www.inf.ufsc.br/~mauro.roisenberg/ine6103/slide/CursoRedesNeurais.pdf`. Acesso em 27 nov. 2022.

SAK, Hasim; SENIOR, Andrew W; BEAUFAYS, Françoise. Long short-term memory recurrent neural network architectures for large scale acoustic modeling, 2014.

SCHIELE, Julian; KOPERNA, Thomas; BRUNNER, Jens O. Predicting intensive care unit bed occupancy for integrated operating room scheduling via neural networks. **Naval Research Logistics (NRL)**, Wiley Online Library, v. 68, n. 1, p. 65–88, 2021.

SCHMIDHUBER, Jürgen; BLOG, A. The 2010s: our decade of deep learning/outlook on the 2020s. **The recent decade's most important developments and industrial applications based on our AI, with an outlook on the 2020s, also addressing privacy and data markets**, 2020.

SEZER, Omer Berat; GUDELEK, Mehmet Ugur; OZBAYOGLU, Ahmet Murat. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. **Applied soft computing**, Elsevier, v. 90, p. 106181, 2020.

SVOZIL, Daniel; KVASNICKA, Vladimir; POSPICHAL, Jiri. Introduction to multi-layer feed-forward neural networks. **Chemometrics and intelligent laboratory systems**, Elsevier, v. 39, n. 1, p. 43–62, 1997.

TAYLOR, Sean J; LETHAM, Benjamin. Forecasting at scale. **The American Statistician**, Taylor & Francis, v. 72, n. 1, p. 37–45, 2018.

TELLO, Manuel *et al.* Machine learning based forecast for the prediction of inpatient bed demand. **BMC medical informatics and decision making**, Springer, v. 22, n. 1, p. 1–13, 2022.

WILLIAMS, Ronald J; ZIPSER, David. Gradient-based learning algorithms for recurrent. **Backpropagation: Theory, architectures, and applications**, v. 433, p. 17, 1995.

YE, Andre. **Recurrent neural Networks Explained and Visualized from the Ground Up**. Disponível em: `https://towardsdatascience.com/recurrent-neural-networks-explained-and-visualized-from-the-ground-up-51c023f2b6fe`. Acesso em: 7 jul. 2023.

Apêndices

APÊNDICE A – ARTIGO

Predição de Ocupação de Leitos Hospitalares de Terapia Intensiva no Curto Prazo Utilizando Redes Neurais Artificiais

João J. Paraquetti[1], **Jerusa Marchi**[1], **André W. Zibetti**[1]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC) - Florianólis, SC – Brasil

joao.jandre@grad.ufsc.br, jerusa.marchi@ufsc.br, andre.zibetti@ufsc.br

Abstract. *Effectively predicting demand for healthcare services is important in virtually all types of public services, as allocating fewer resources than necessary can result in a loss of quality in consumer care, while allocating excessive resources can lead to wastage. In the specific case of hospitals, excessive waiting times for care can lead to patient mortality, and the high demand for public hospital services necessitates optimal resource utilization to avoid waste. This study provides a brief overview of the problem of predicting hospital bed occupancy and, using the available data from DATASUS on hospitals in the state of Santa Catarina, decomposes the time series of bed occupancy to explain its behavior. The obtained and processed data were used to train various LSTM networks with the aim of predicting ICU bed occupancy in the state of Santa Catarina. Finally, the study compares the developed models, concluding that the best results were achieved with the model from the third government of the period studied, using only 4 neurons in the hidden layer and considering the previous 3 weeks for prediction.*

Resumo. *Prever com eficácia a demanda por atendimentos é importante em praticamente todos os tipos de serviços ofertados ao público, pois alocar menos recursos do que o necessário pode levar a perda da qualidade do atendimento ao consumidor e alocar recursos a mais pode significar desperdício de tais recursos. No caso específico de hospitais, a espera demasiada por atendimento pode levar ao óbito de pacientes e a alta demanda por serviços hospitalares públicos obriga um uso ótimo de recursos, evitando desperdício. Este trabalho dá uma breve visão sobre o problema de predição de ocupação de leitos hospitalares, e usando os dados disponíveis pelo DATASUS dos hospitais do estado de Santa Catarina, faz uma decomposição da série temporal da ocupação de leitos, buscando explicar comportamento das mesmas, os dados obtidos e tratados foram usados para treinamento de diversas redes LSTM, com o objetivo de prever a ocupação de leitos de UTI no estado de Santa Catarina. Finalmente, o trabalho faz uma comparação entre os modelos desenvolvidos, concluindo que a rede que apresentou os melhores resultados foram obtidos com o modelo do terceiro governo do período estudado, utilizando apenas 4 neurônios na camada oculta, e considerando 3 semanas anteriores para a predição.*

1. Introdução

Projetar e implementar políticas eficazes de gestão de capacidade hospitalar e decisões de alocação de profissionais é um desafio crítico em todos os sistemas de saúde. Um

desencontro entre número de leitos disponíveis e a demanda, assim como a quantidade de profissionais de saúde alocados podem afetar negativamente diversos indicadores de performance dos hospitais, como tempo de espera, tamanho da fila de espera, qualidade do atendimento, assim como a satisfação dos pacientes e profissionais de saúde [Tello et al. 2022].

Levando isso em conta, existe uma necessidade amplamente reconhecida de prever a ocupação de leitos de hospital. Quanto melhores previsões pudermos fazer, mais eficientemente poderemos planejar com antecedência e, como resultado, o uso de recursos é otimizado e melhores cuidados podem ser prestados aos pacientes [Kutafina et al. 2019].

O problema de previsão de ocupação de leitos já é explorado há décadas, com diversas soluções propostas. Classicamente, as abordagens numéricas eram populares, diversos autores propunham o uso de regressões, como em [Beenhakker 1963], muitas vezes utilizando-se de séries temporais, como proposto em [Littig and Isken 2007]. O modelo auto-regressivo integrado de médias móveis, ou ARIMA, também foi outra técnica explorada por autores, como em [Earnest et al. 2005] onde foi feito um estudo retrospectivo utilizando esse modelo para prever o número de leitos ocupados durante o surto de SARS de 2003 em Singapura.

Nas últimas décadas, abordagens orientadas a dados e aprendizado de máquina provaram sua eficiência para tarefas de previsão. No entanto, apenas um progresso limitado foi feito na aplicação desses métodos à previsão de ocupação de leitos hospitalares [Kutafina et al. 2019]. Entretanto, nos últimos anos, houve um crescimento do uso de técnicas de aprendizado de máquina para previsão de leitos de UTI, especialmente durante a pandemia de COVID-19.

No Brasil, todos os dados coletados por hospitais públicos são armazenados pelo ministério da saúde [Carvalho 2009], de forma que, técnicas capazes de processar grandes quantidades de dados poderiam usar essas informações para construir modelos precisos, como é o caso das técnicas de aprendizado de máquina. Este trabalho visa aplicar redes neurais do tipo Long Short Term Memory para previsão de ocupação de leitos hospitalares em curto prazo.

2. Fundamentação teórica

2.1. Sistema de Informações Hospitalares-SIH

O Sistema de Informações Hospitalares, ou SIH, teve seu início na década de 1970, foi implantado com a intenção de controlar o pagamento dos serviços prestados pelos hospitais contratados [Lessa et al. 2000]. Até 1991, o sistema passou por diversos nomes, para então finalmente ser renomeado para Sistema de Informações Hospitalares [Levcovitz and Pereira 1993]. Todo o acervo de informações e valores do sistema antigo passou a compor a base do SIH/SUS[Lessa et al. 2000].

O sistema transcreve todos os atendimentos provenientes de internações hospitalares que foram financiadas pelo SUS, e após o processamento, gera relatórios para os gestores, possibilitando fazer os pagamentos dos estabelecimentos de saúde. Além disso, o nível federal recebe mensalmente uma base de dados de todas as internações autorizadas (aprovadas ou não para pagamento), para que possam ser repassados às Secretarias de Saúde [IBGE b].

A AIH é o documento hábil para identificar o paciente e os serviços prestados sob regime de internação hospitalar pelo SUS. Fornece informações para o gerenciamento do sistema e, através dele, os hospitais, profissionais e serviços auxiliares de diagnóstico e terapia receberão pelos serviços prestados ao usuário [Lessa et al. 2000]. A AIH pode ser separada em 5 seções: identificação do hospital, caracterização da internação, procedimentos especiais, serviços profissionais e caracterização da assistência prestada. Para os objetivos do trabalho, apenas as seções de caracterização da internação e procedimentos especiais serão relevantes, pois é onde estão descritas as informações sobre o paciente e informações sobre internações sobre o uso da unidade de tratamento intensivo, UTI. Uma descrição completa dos dados das AIH pode ser encontrada no anexo 1 de [Lessa et al. 2000].

Os dados disponibilizados livremente pelo DATASUS são um conjunto reduzido e anonimizado das AIH, de forma que não é possível identificar os pacientes de nenhuma forma, o que poderia ser um fator limitante para certos tipos de estudos. Entretanto como este trabalho visa apenas estudar o comportamento da ocupação de leitos de UTI, não é necessária a identificação de pacientes.

2.2. Séries Temporais

Uma série temporal é muitas vezes o resultado da observação de algum processo, onde valores são coletados a partir de medições feitas em instantes de tempo uniformemente espaçados, de acordo com uma determinada taxa de amostragem. Uma série temporal pode, portanto, ser definida como um conjunto de instantes de tempo contíguos [Esling and Agon 2012].

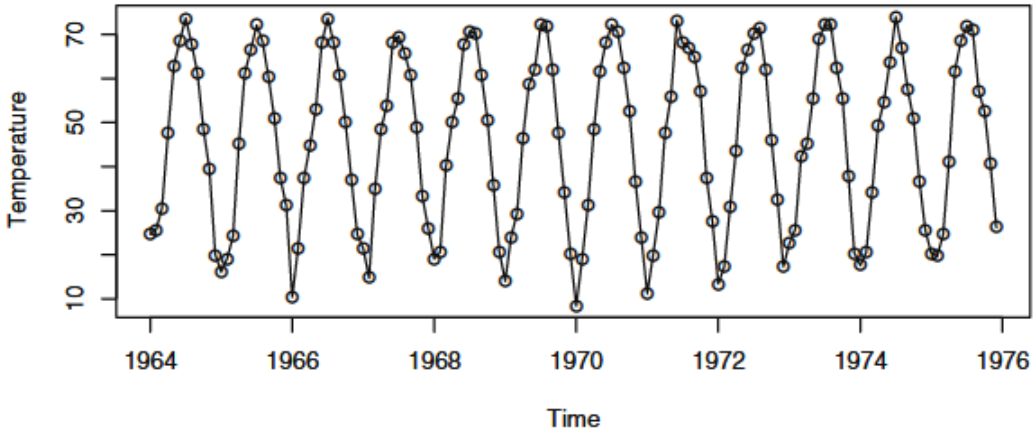


Figura 1. Temperaturas médias mensais, Dubuque, Iowa

Uma série temporal discreta é aquela em que o conjunto T_0 de tempos nos quais as observações são feitas é um conjunto discreto, como por exemplo na figura 1, quando as observações são feitas em pontos fixos [1]. Séries temporais contínuas são obtidas quando as observações são registradas continuamente durante algum intervalo de tempo, por exemplo, quando T_0 = [0, 1] [Brockwell and Davis 2002].

[1]Neste gráfico, as observações são representadas pelos círculos no gráfico, enquanto que a curva traçada é apenas uma aproximação dos dados da série seriam caso a mesma fosse contínua

Existem algumas características comuns que são observadas ao trabalhar com séries temporais, como tendência, sazonalidade, ciclos e estacionariedade. Uma tendência existe quando há um aumento ou diminuição de longo prazo nos dados, não necessariamente linear [Hyndman and Athanasopoulos 2018]. Um padrão sazonal ocorre quando uma série temporal é afetada por fatores sazonais, como a época do ano ou o dia da semana. A sazonalidade é sempre de um período fixo e conhecido [Hyndman and Athanasopoulos 2018]. Um ciclo ocorre quando os dados exibem aumentos e quedas que não são de uma frequência fixa. A duração dessas flutuações é geralmente de pelo menos 2 anos [Hyndman and Athanasopoulos 2018]. Uma série temporal é dita estacionaria se não possui uma mudança sistemática na média, ou seja, sem tendência, se não possui nenhuma mudança sistemática na variância, e se variações estritamente periódicas tenham sido removidas [Chatfield 2013].

Uma série pode ser univariada, ou seja, quando uma única variável é observada ao longo do tempo; ou multivariada, quando várias séries abrangem simultaneamente várias dimensões dentro do mesmo intervalo de tempo [Esling and Agon 2012]. Uma série temporal pode cobrir todo o conjunto de dados fornecidos pela observação de um processo, e pode ter um comprimento considerável. Além disso, elas são consideradas suaves, isto é, valores subsequentes estão dentro de faixas predizíveis uns dos outros [Esling and Agon 2012].

2.3. Redes Neurais Long Short Term Memory

As redes *Long short term memory*, ou LSTM, são do tipo recorrentes, ou seja, possuem retroalimentação, o que cria possibilidade da rede ter uma memória de curto prazo. Entretanto, aprender a armazenar informações em intervalos de tempo prolongados por meio de retropropagação recorrente leva muito tempo, principalmente devido ao fluxo de retorno de erro insuficiente e decadente [Hochreiter and Schmidhuber 1997]. Com métodos convencionais de otimizar redes neurais recorrentes, como *Real time recurrent learning* [Robinson and Fallside 1987] e *Back-propagation through time* [Williams and Zipser 1995] sinais de erro "retrocedendo no tempo" tendem a explodir, ou sumir, com o primeiro caso levando a pesos oscilantes, e o segundo levando a tempos de treinamento muito longos ou simplesmente não funcionando [Hochreiter and Schmidhuber 1997].

Esse tipo de rede é especialmente efetiva para fazer predições baseadas em séries temporais, já que ela consegue capturar atrasos de grande duração entre os eventos. Ao contrário das redes neurais convencionais, que são compostas por neurônios ou nodos, redes LSTM geralmente são compostas por células, que possuem um estado interno[2], portões de entrada, saída e de esquecimento. Um exemplo de uma unidade deste tipo é mostrada na figura 2.

A função do *forget gate* é decidir quais informações não são mais úteis no estado da célula, ele é alimentado com a saída anterior da célula e a entrada atual, as entradas são multiplicadas por uma matriz de pesos e então adicionadas de um viés [Academy]. O resultado é passado por uma função de ativação que fornece uma saída binária, geralmente a sigmoide logística [Graves 2012]. Se para um determinado estado de célula a saída for 0, a informação é esquecida, se a saída for 1, a informação é retida para uso futuro.

[2]Também conhecido como CEC ou *Constant Error Carousel*

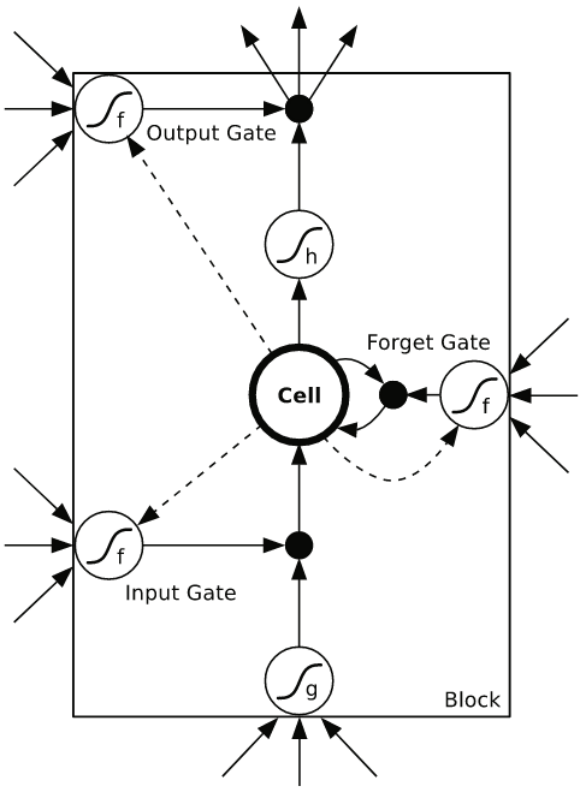


Figura 2. Arquitetura de uma célula LSTM

A regulagem de quais novas informações serão adicionadas ao estado da célula é feita através do *input gate*. Similarmente ao *forget gate*, o *input gate* recebe a saída anterior da célula e a entrada atual, multiplica as mesmas por uma matriz de pesos, adicionando um viés e finalmente passando por uma função de ativação que fornece uma saída binária, como a sigmoide logística. Essa saída então é multiplicada pela saída da função f_g, fazendo com que o resultado da mesma seja adicionado ou não ao estado interno da célula [Graves 2012]. O *output gate* segue a mesma lógica do *input gate*, porém com o objetivo de decidir o quanto o estado interno da célula vai influenciar na saída.

Nenhuma função de ativação é aplicada dentro da célula. A função de ativação f_f dos portões é geralmente a sigmoide logística, de modo que as ativações dos portões estejam entre 0 (portão fechado) e 1 (portão aberto). As funções de ativação de entrada e saída da célula (f_g e f_h) são geralmente a tangente hiperbólica ou a sigmoide logística, embora em alguns casos f_h seja a função de identidade [Graves 2012].

As linhas tracejadas representam conexões internas, que permitem que os portões da célula tenham acesso ao estado interno da célula, mesmo quando o portão de saída está fechado, essas conexões são chamadas de *peepholes*. Os *peepholes* não estavam presentes na arquitetura original das LSTMs, tendo sido propostos para melhor a performance em tarefas altamente não lineares, permitindo que a célula aprenda a medir intervalos de tempo precisos [Gers et al. 2002]. Nenhuma das outras conexões dentro do bloco são ponderadas. As únicas saídas do bloco para o resto da rede emanam da multiplicação da porta de saída, ou seja, de certo modo, cada célula se comporta como um perceptron, recebendo n entradas porém emanando apenas uma saída.

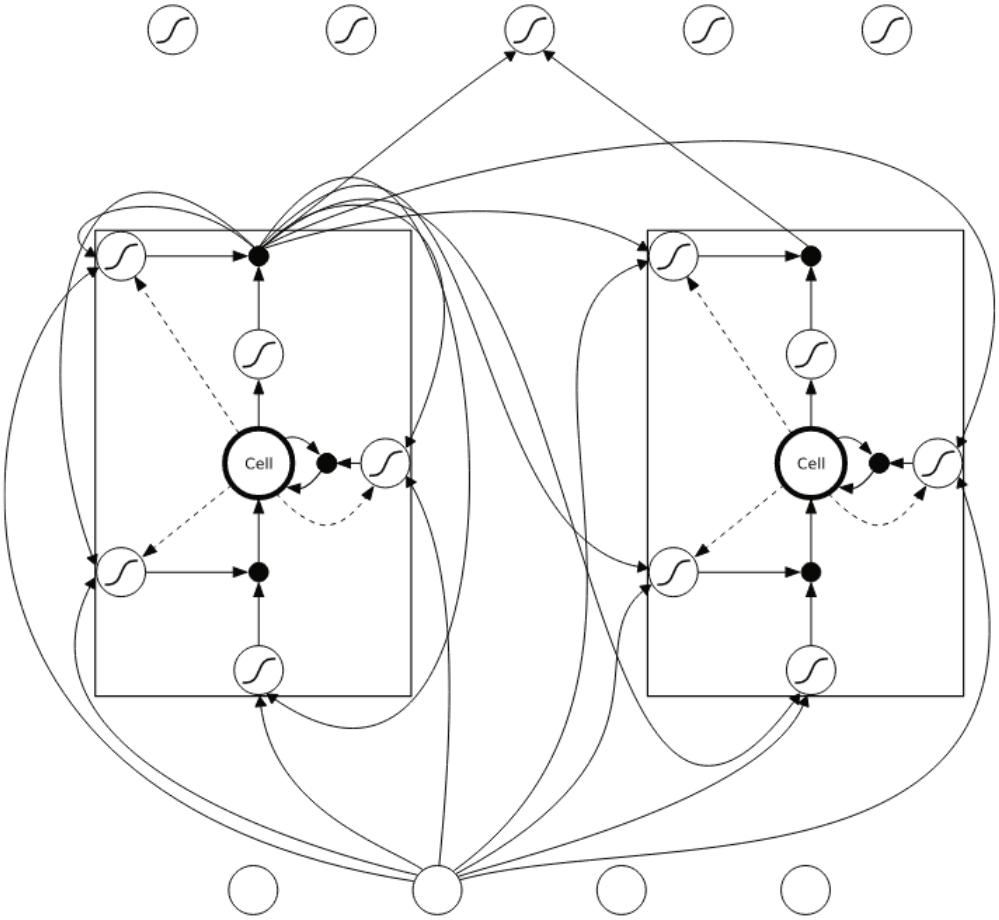


Figura 3. Rede LSTM

Um exemplo de rede LSTM pode ser observado na figura 3, a rede demonstrada possui quatro neurônios de entrada, uma camada escondida com duas células LSTM e cinco neurônios de saída. Para não poluir a imagem, nem todas as conexões são mostradas. É importante notar que cada célula tem quatro entradas, porém apenas uma saída, também é interessante notar as conexões recorrentes das células.

3. Trabalhos Correlatos

Existem diversas propostas, descritas na literatura, de algoritmos de predição de leitos hospitalares, sendo que a maioria das soluções mais atuais focam na utilização do aprendizado de máquina para resolver essa tarefa. Foi feita uma revisão sistemática da literatura existente, com objetivo de encontrar os trabalhos mais relevantes sobre predição de ocupação de leitos de UTI utilizando aprendizado de máquina. A tabela 1 ilustra um resumo das características principais dos trabalhos encontrados, que são mais correlacionados a este.

Tabela 1. Comparação dos Trabalhos Correlatos

Título	Características
Predicting intensive care unit bed occupancy for integrated operating room scheduling via neural networks	Utilizam informações extremamente detalhadas sobre os pacientes e sua estadia para prever a ocupação de leitos, utilizando RNAs.
COVID-19 ICU demand forecasting: A two-stage Prophet-LSTM approach	Utilizam dois tipos de ML, Prophet e LSTM, para previsão da ocupação de leitos durante a pandemia de COVID-19.
Artificial neural networks for short-term forecasting of cases, deaths, and hospital beds occupancy in the COVID-19 pandemic at the Brazilian Amazon	Utilizam RNAs para prever casos de COVID-19, mortes e ocupação de leitos no estado do Pará.

4. Preparação dos Dados

Como disposto na seção 2.1, o DATASUS disponibiliza livremente os dados reduzidos das AIH. Para o escopo desse trabalho, apenas algumas informações foram consideradas importantes, a data de hospitalização, a data de alta, o número de dias na UTI e o tipo de UTI. Como as datas exatas de entrada e saída de UTI dos pacientes não estão disponíveis, não é possível determinar com uma precisão diária quantos pacientes estavam ocupando leitos de UTI, portanto, foi necessário utilizar alguma estratégia para arbitrariamente decidir quando considerar que alguém estava ocupando um leito.

A estratégia decidida foi considerar que o paciente é admitido na UTI no mesmo dia que é hospitalizado, e que permanece nela pelo número de dias associados a estadia na UTI, quando então recebe alta ou é transferido para um leito comum, conforme os dados de internação. Essa perda de precisão fez com que a estratégia de agregação de dados diária fosse deixada de lado em favor de uma estratégia de agregação de dados com granularidade de semanas. A estratégia semanal foi escolhida pois a média de tempo de internação dos pacientes em leitos de UTI é de aproximadamente 6,7 dias, enquanto que a mediana é de 4 dias, ou seja, grande parte das pessoas ocupa um leito de UTI por quase uma semana. Entretanto, uma granularidade maior dos dados ainda seria benéfica.

O pré-processamento dos dados para análise preliminar consistiu dos seguintes passos:

1. Selecionar apenas as colunas de interesse;
2. Transformar as datas do tipo string para o formato datetime do Pandas;
3. Criar uma nova coluna para armazenar a data de alta da UTI;
4. Calcular a data de saída da UTI, utilizando a data da hospitalização como data de entrada na UTI e somando o número de dias de estadia na UTI (diminuindo de um, já que o dia que o paciente foi hospitalizado também entra na conta);

5. Agrupar os registros por semanas, sendo que é considerado que se a pessoa estava qualquer quantidade de tempo na UTI em uma dada semana, ela estava ocupando um leito de UTI naquela semana;
6. Contar a quantidade de registros em cada semana e criar uma nova tabela apenas com o número da semana e a quantidade de leitos ocupados.

5. Análise preliminar dos dados

A figura 4 mostra a ocupação de leitos de UTI semanal entre a primeira semana do ano de 2010 e a última semana do ano de 2018, já é possível verificar visualmente que existe uma certa tendência de aumento do número da ocupação dos leitos de UTI, assim como uma certa sazonalidade, com um aumento no número de casos no inverno, e uma diminuição no verão.

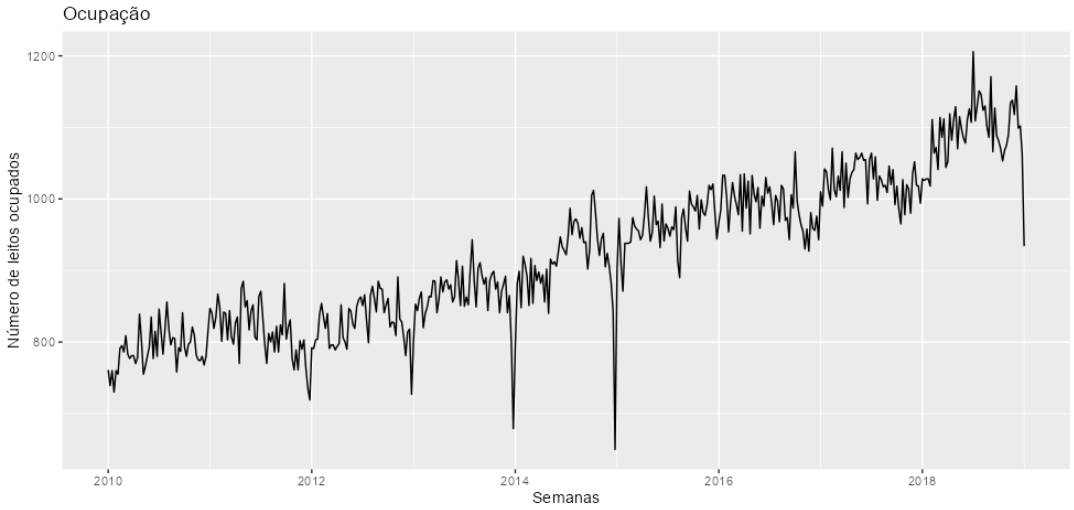


Figura 4. Ocupação de leitos de UTI semanal entre 2010 e 2018

O gráfico mostrado na figura 5 apresenta uma comparação da ocupação semanal de leitos de UTI entre os anos analisados, neste gráfico, valores mais próximos do centro são menores, enquanto que valores mais longe são maiores. A figura demonstra uma similaridade entre os comportamentos dos anos, além disso, ela reforça a noção de que cada ano passado existe um aumento na ocupação de leitos, explicitado pelo aumento da circunferência dos círculos a cada ano. Para tentar explicar esse comportamento, foram buscados dados do Instituto Brasileiro de Geografia e Estatística sobre a população do estado de Santa Catarina de 2010, assim como sua projeção para os anos seguintes [IBGE a].

Usando esses dados, dois gráficos foram produzidos: o gráfico A, com os números absolutos de ocupação de leitos para cada ano; o gráfico B, com o número de ocupação de leitos em relação à estimativa populacional daquele ano. A figura 6 apresenta a comparação dos gráficos, pode-se observar que apesar de existir um pequeno achatamento da curva, o crescimento da ocupação de leitos ainda é extremamente evidente, e portanto, o crescimento populacional não explica completamente o aumento paulatino da ocupação de leitos.

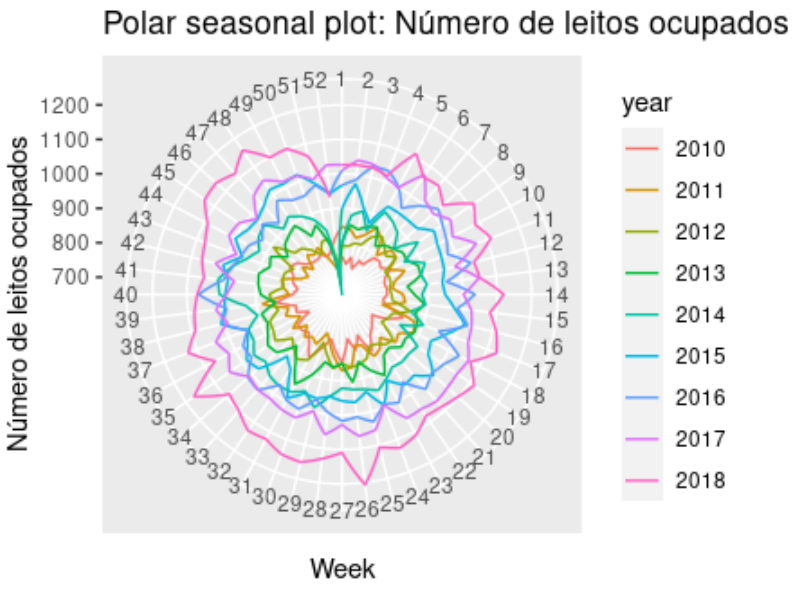


Figura 5. Comparação da ocupação de leitos de UTI semanal entre 2010 e 2018

Também foi verificado se este aumento da demanda por leitos de UTI acompanha um aumento da demanda por hospitalizações em geral, entretanto, como pode-se observar na figura 7, enquanto que a demanda por hospitalizações cresceu em torno de 22%, a demanda por leitos de UTI cresceu 35%. É provável que outros fatores externos causem essa tendência de crescimento da demanda por leitos de UTI, como aumento da oferta de leitos ou envelhecimento da população, entretanto, uma análise mais aprofundada sobre essa causa foge do escopo deste trabalho.

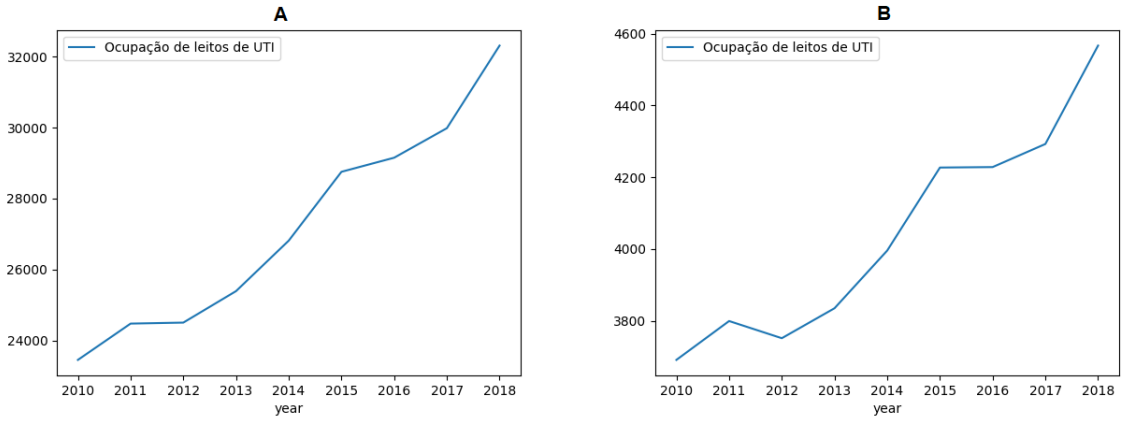


Figura 6. Comparação entre números de ocupação de leitos de UTI anuais absolutos, e relativos ao crescimento populacional

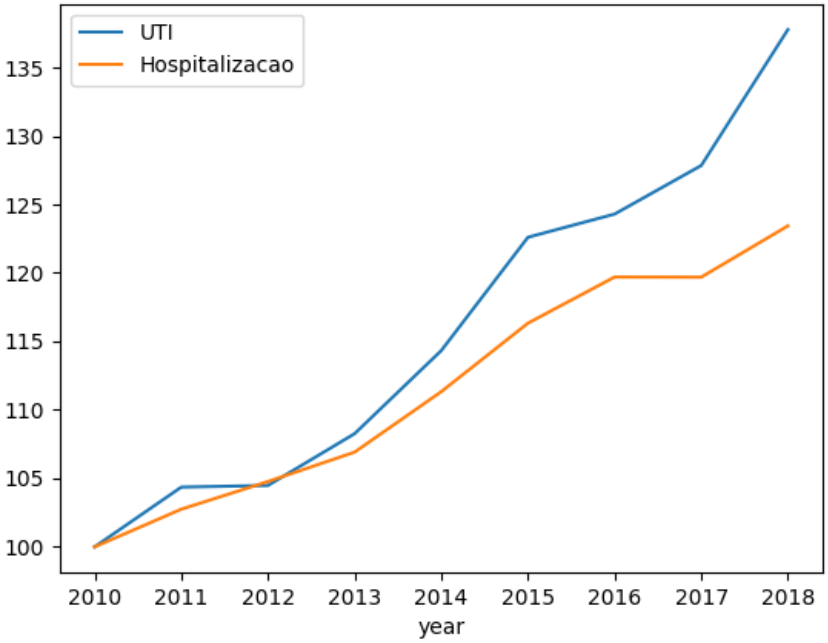


Figura 7. Comparação do crescimento da demanda por leitos de UTI e da demanda de hospitalizações

Finalmente, foi feita uma decomposição da série temporal, utilizando o método de decomposição de Loess, baseada na função de suavização local ponderada por polinômios [Cleveland et al. 1990], a mesma encontra-se na figura 8. A decomposição da série temporal está dividida em três painéis (tendência, sazonalidade e resíduo), cada um com uma componente da série. Esses componentes podem ser somados para reconstruir os dados mostrados no painel superior (data). Observa-se que o componente sazonal não apresenta mudança significativa ao longo do tempo, de modo que qualquer ano consecutivo apresenta padrão semelhante. O componente de resíduo mostrado no painel inferior é

o que resta quando os componentes sazonal e de tendência-ciclo são subtraídos dos dados. As barras cinzas à esquerda de cada painel mostram as escalas relativas dos componentes. Cada barra cinza representa o mesmo comprimento, mas devido às diferentes escalas dos gráficos, as barras variam em tamanho. Analisando a figura, é possível confirmar que existe uma tendência clara de crescimento da série, assim como uma sazonalidade evidente, que tende a levar à um aumento dos valores da série no meio do ano.

A análise dos dados serviu para confirmar que a série tratada tem um comportamento dentro do esperado de uma série temporal comum, isto é, possui uma sazonalidade clara, além de uma tendencia de crescimento quase constante, com resíduos relativamente pequenos, ou seja, todas essas características mostram que a série não segue um padrão aleatório, mas possuí um mecanismo por trás que pode eventualmente ser aproximado por alguma técnica, como redes neurais, de maneira que possamos atingir predições com alta acurácia. Ademais, é interessante que a análise seja feita antes do início do treinamento das redes, pois desta forma, podemos criar uma intuição de quais seriam as melhores maneiras de treinar a rede. Por exemplo, como a autocorrelação da série é relativamente alta, podemos tentar utilizar desde o início mais de uma semana como entrada da rede.

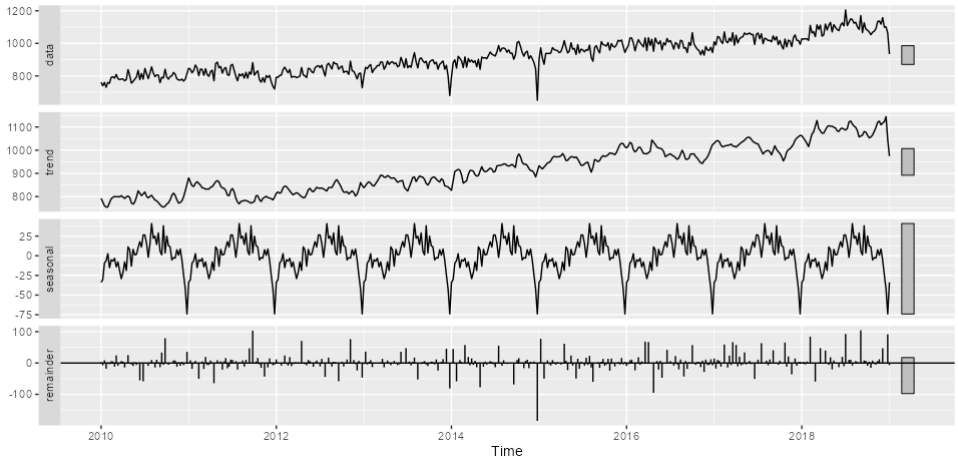


Figura 8. Decomposição da série temporal

6. Treinamento das redes LSTM

Antes de iniciar o treinamento de redes neurais é necessário fazer um tratamento dos dados, afim de extrair a melhor performance das mesmas. O primeiro passo do tratamento foi a remoção dos *outliers* estatísticos, para isso foi feito o cálculo dos quantis de 1% e 99% dos dados, valores fora dessa faixa foram descartados.

O próximo passo foi transformar a série temporal em estacionária, removendo a tendência da mesma, um procedimento comum para isso é tomar a diferença da série[Coghlan 2015], isto é, cada valor é subtraído do anterior, de forma que a série temporal resultante representa as diferenças entre os valores originais. O resultado desse processo pode ser observado na figura 9. Ademais, os dados foram normalizados entre -1 e 1, outro procedimento comum para o treinamento de redes neurais, a normalização foi feita com auxílio da biblioteca *open source* scikit-learn do Python.

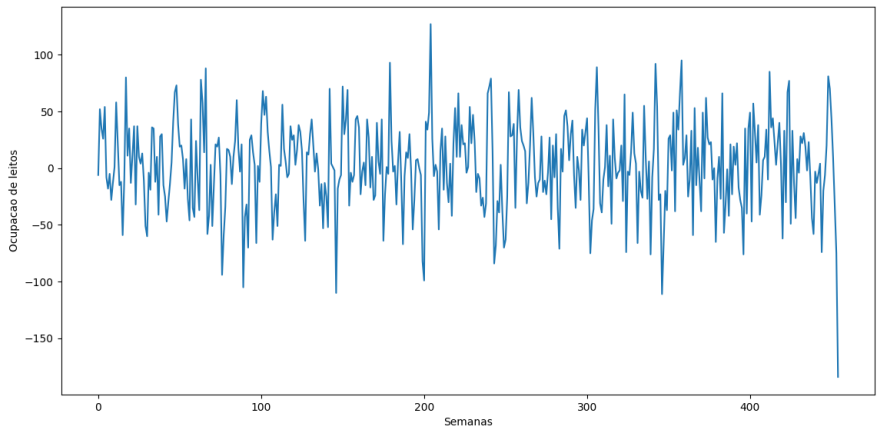


Figura 9. Série estacionária

Para que a rede treinada utilize mais do que apenas uma semana como entrada para predizer a próxima, ainda é necessário fazer mais uma última transformação dos dados de entrada. Precisamos modificar a dimensão do vetor de entrada da rede, para que o mesmo possua a quantidade correta de semanas por tupla. Por exemplo, para que seja possível utilizar as últimas três semanas para predizer a próxima, é necessário que cada tupla do vetor de entrada possua três valores, um com o valor da ocupação no tempo $T - 2$ outro no tempo $T - 1$ e finalmente um no tempo T. Na biblioteca Pandas, primeiramente a quantidade de colunas necessária é criada no *dataframe* de entrada, então cada coluna é populada com os valores necessários.

A função de ativação utilizada foi a tangente hiperbólica, a função recorrente[3] utilizada foi a sigmoide logística, ambas as funções são comumente usadas em redes LSTM, como apresentado na seção 2.3. O otimizador escolhido foi o Adam, pois o mesmo é robusto e muito utilizado em uma gama de problemas de otimização na área de aprendizado de máquina[Kingma and Ba 2014]. A função de perda utilizada foi o erro médio quadrático. Todas as redes foram treinadas com 300 épocas, utilizando *early stopping* com paciência de 6, de forma que nenhuma das redes chegou a realmente treinar pelo número máximo de épocas. O conjunto de treinamento sempre representava 75% dos dados, enquanto que o conjunto de testes era composto pelo restante. Como as redes LSTM possuem um estado interno, entre o fim do treinamento e o início dos testes, as redes foram alimentadas com os dados de treinamento uma última vez, para que a rede possuísse um estado interno construído antes de começar os testes.

A medida de acurácia escolhida utilizada para julgar a qualidade das redes treinadas foi a raiz do erro quadrático médio, ou RMSE. Como o efeito de cada erro no RMSE é proporcional ao erro quadrático, erros maiores tem um efeito desproporcionalmente grande no RMSE, fazendo com que o RMSE seja mais sensível à *outliers*, para o objeto de estudo deste trabalho, é extremamente importante que mesmo os *outliers* sejam previstos com uma acurácia decente, portanto a escolha da medida de acurácia se justifica. Além disso, o RMSE sempre está na mesma escala dos dados, portanto serve apenas para

[3] A função de ativação dos portões de entrada, saída e esquecimento das células.

comparações dentro do mesmo conjunto de dados. Historicamente o RMSE sempre foi uma medida popular, principalmente por causa da sua relevância teórica na modelagem estatística [Hyndman and Koehler 2006]. A métrica secundária de comparação utilizada foi o MAPE, ou erro percentual absoluto médio, utilizado na comparação entre os modelos desenvolvidos.

Foram criados diversos modelos com intenção de buscar qual tipo de modelagem levaria à uma acurácia melhor das redes LSTM. Além disso, para cada modelo, os hiper-parâmetros foram ajustados manualmente na tentativa de melhorar o desempenho das redes. O primeiro modelo foi o modelo "generalista", que utiliza todos os dados disponíveis para treinamento e validação. Os demais trabalharam apenas com uma parte dos dados, com a intenção explorar possíveis similaridades, foram criados modelos treinados separados por períodos de governos de presidentes da república, além de modelos treinados apenas por certos períodos do ano, como outono e inverno.

6.1. Modelo generalista

Como apresentado anteriormente, o primeiro modelo treinado utiliza indiscriminadamente todos os dados disponíveis, ou seja, os dados de 2010 até 2016 foram utilizados para treinamento da rede, enquanto que os dados de 2017 até 2018 foram utilizados para a validação da mesma. A tabela 2 apresenta os resultados das redes que tiveram maior desempenho, na primeira coluna é mostrada a quantidade de semanas antecedentes utilizadas para prever a semana seguinte, a segunda coluna apresenta a arquitetura da rede[4] e a terceira coluna o RMSE.

Tabela 2. Modelo generalista

Semanas consideradas	Arquitetura da rede	RMSE alcançado
4 semanas anteriores	4 neurônios	46,928
3 semanas anteriores	4 neurônios	46,977
3 semanas anteriores	3 neurônios	48,794
2 semanas anteriores	4 neurônios	51,154
4 semanas anteriores	**4/2 neurônios**	**43,408**
3 semanas anteriores	4/2 neurônios	48,209
2 semanas anteriores	4/2 neurônios	52,529

É possível ver que os RMSEs alcançados são relativamente próximos, entretanto, redes que consideram uma quantidade de semanas maior que dois performam relativamente melhor, com a melhor rede considerando quatro semanas anteriores e possuindo duas camadas com quatro e dois neurônios, respectivamente. Durante o treinamento, redes que consideravam uma quantidade maior que quatro semanas foram treinadas, porém após a quarta semana, o desempenho das redes começa a cair significativamente, o que indica que a autocorrelação da série temporal tende a cair consideravelmente após um atraso de quatro semanas.

[4]Quando a rede possui mais de uma *hidden layer* as camadas são representadas como X/Y, onde X e Y são a quantidade de neurônios nas camadas um e dois, respectivamente.

6.2. Modelos separados por governos federais

Os próximos três modelos criados segregam os dados nos respectivos períodos de atuação dos presidentes da república, a ideia por trás dessa separação é que cada governo tende a ter certas políticas públicas sobre a saúde, o que poderia portanto impactar o comportamento da série temporal, de maneira que, redes treinadas especificamente para um certo governo poderiam ter um desempenho mais satisfatório.

O primeiro modelo treinado considera os primeiros quatro anos do período dos dados coletados, refletindo o período do primeiro governo da então presidente da república Dilma Rousseff, o segundo modelo reflete o período do segundo governo da ex-presidente, até seu eventual término via impeachment, o último modelo dentre os três reflete o período de governo do então presidente Michel Temer.

Tabela 3. Modelo do primeiro governo

Semanas consideradas	Arquitetura da rede	RMSE alcançado
4 semanas anteriores	4 neurônios	47,107
3 semanas anteriores	4 neurônios	42,311
3 semanas anteriores	3 neurônios	53,142
2 semanas anteriores	4 neurônios	44,498
4 semanas anteriores	4/2 neurônios	45,891
3 semanas anteriores	4/2 neurônios	43,072
2 semanas anteriores	**4/2 neurônios**	**41,570**

A tabela 3 apresenta os melhores resultados obtidos no treinamento das redes que seguem o modelo do primeiro governo, é possível perceber que houve uma melhora na acurácia das redes treinadas em relação ao modelo genérico, possivelmente pois a conjectura levantada anteriormente realmente tem uma certa validade, ou seja, é provável que redes treinadas considerando as políticas específicas de certo governo tenham uma acurácia melhor na previsão da ocupação de leitos.

Tabela 4. Modelo do segundo governo

Semanas consideradas	Arquitetura da rede	RMSE alcançado
4 semanas anteriores	4 neurônios	62,599
3 semanas anteriores	**4 neurônios**	**50,663**
3 semanas anteriores	3 neurônios	52,195
2 semanas anteriores	4 neurônios	64,488
4 semanas anteriores	4/2 neurônios	61,967
3 semanas anteriores	4/2 neurônios	51,170
2 semanas anteriores	4/2 neurônios	61,633

Os resultados obtidos para as redes treinadas no segundo governo do período, apresentados na tabela 4, foram os piores dentre todos os modelos, uma possível explicação é que a instabilidade política do período tenha afetado o comportamento da

série, fazendo com que a mesma seja de mais difícil interpretação e previsão pelas redes LSTM.

Tabela 5. Modelo do terceiro governo

Semanas consideradas	Arquitetura da rede	RMSE alcançado
4 semanas anteriores	4 neurônios	41,745
3 semanas anteriores	**4 neurônios**	**37,038**
3 semanas anteriores	3 neurônios	40,456
2 semanas anteriores	4 neurônios	40,967
4 semanas anteriores	4/2 neurônios	42,603
3 semanas anteriores	4/2 neurônios	39,411
2 semanas anteriores	4/2 neurônios	40,452

A tabela 5 apresenta os melhores resultados obtidos nos treinamentos das redes seguindo o modelo do terceiro governo, dentre todos os modelos feitos, as redes deste modelo foram as que previram melhor a ocupação de leitos de UTI, segundo a métrica adotada. Isto pode ser dado devido a uma combinação dos fatores anteriormente mencionados, ou seja, as políticas de governo na saúde estabilizaram a série, além disso, a instabilidade política foi diminuída.

6.3. Modelos separados por estações do ano

Como foi apresentado na seção 5, a série temporal que é objeto de estudo deste trabalho possui uma sazonalidade significativa, levando isso em conta, os últimos dois modelos criados focam em separar o ano em duas partes, outono/inverno e primavera/verão, com a intenção de treinar redes que são especializadas em predizer o comportamento da ocupação de leitos durante certas estações.

Tabela 6. Modelo primavera/verão

Semanas consideradas	Arquitetura da rede	RMSE alcançado
4 semanas anteriores	4 neurônios	53,827
3 semanas anteriores	**4 neurônios**	**50,606**
3 semanas anteriores	3 neurônios	51,004
2 semanas anteriores	4 neurônios	52,202
4 semanas anteriores	4/2 neurônios	51,302
3 semanas anteriores	4/2 neurônios	52,189
2 semanas anteriores	4/2 neurônios	54,166

Analisando a tabela 6, percebe-se que os resultados são relativamente ruins para todas as redes treinadas, parte do motivo pode ser o fato de que em vários dos anos estudados, existe uma oscilação grande no número registrado de pacientes ocupando leitos na última semana do ano, que pode afetar negativamente o treino da rede. Outra possibilidade é que a sazonalidade no período da primavera e verão é pouco pronunciada, e portanto de mais difícil interpretação pelas redes neurais.

Tabela 7. Modelo outono/inverno

Semanas consideradas	Arquitetura da rede	RMSE alcançado
4 semanas anteriores	4 neurônios	46,246
3 semanas anteriores	**4 neurônios**	**42,377**
3 semanas anteriores	3 neurônios	52,799
2 semanas anteriores	4 neurônios	54,168
4 semanas anteriores	4/2 neurônios	52,257
3 semanas anteriores	4/2 neurônios	45,362
2 semanas anteriores	4/2 neurônios	52,884

Finalmente, analisando a tabela 7, percebe-se que os resultados são relativamente parecidos com os do modelo genérico, levando em conta isso e os resultados do modelo da primavera e verão, pode-se afirmar que modelos que levam em conta apenas estações específicas do ano não parecem ter um desempenho significativamente melhor do que modelos que ignoram isso.

6.4. Comparação entre as redes desenvolvidas

A tabela 8 demonstra uma comparação entre as melhores redes desenvolvidas, além de trazer o RMSE das redes, o MAPE também é mostrado. É possível ver que a rede do terceiro governo se saiu muito melhor, mesmo na métrica secundária adotada, o MAPE, alcançando um erro percentual de 2,434%.

Tabela 8. Melhores redes de cada agrupamento

Agrupamento	Semanas	Arquitetura da rede	RMSE	MAPE
Generalista	4 semanas	4/2 neurônios	43,408	3,392
Primeiro Governo	2 semanas	4/2 neurônios	41,570	3,836
Segundo Governo	3 semanas	4 neurônios	50,663	4,090
Terceiro Governo	**3 semanas**	**4 neurônios**	**37,038**	**2,434**
Primavera/Verão	3 semanas	4 neurônios	50,606	4,170
Outono/Inverno	3 semanas	4 neurônios	42,377	3,432

7. Conclusão

Este trabalho apresentou um estudo preliminar sobre o comportamento da ocupação de leitos hospitalares de UTI no estado de Santa Catarina, explicando o comportamento da série temporal de ocupações de leitos, como visto na seção 5, a ocupação de leitos segue uma tendência clara de crescimento, que não é explicada apenas pelo crescimento populacional, além disso, existe um componente sazonal importante que influencia o comportamento da série. A causa da tendência de crescimento continua desconhecida, e portanto, estudos futuros podem buscar descobrir qual a fonte dessa tendência.

Além disso, este documento apresentou o uso de rede do tipo Long Short Term Memory para o problema de predição de ocupação de leitos, diversas configurações foram feitas, entre elas, foram utilizados de 3 a 6 neurônios, até duas camadas ocultas e

considerando de 2 a 4 semanas anteriores para previsão, além disso, diversas variações no conjunto de dados utilizados foram feitas, primeiramente criando um modelo generalista, com todos os dados, e posteriormente, separando os dados por governos e sazonalidade. Foram apresentadas as motivações para cada modelo de rede proposta, assim como comparações entre os modelos, com o objetivo de buscar quais modelos foram os mais bem sucedidos.

Desta forma, com base nos dados disponíveis, é possível chegar a conclusão de que enquanto um modelo genérico que ignora as nuances dos dados pode ter uma boa acurácia, modelos construídos para situações mais específicas, como por exemplo modelos que consideram apenas um governo, tendem a apresentar desempenhos melhores. O modelo mais bem sucedido foi o modelo que trabalhou sobre os dados do terceiro governo do período, possuindo apenas uma camada oculta com 4 unidades LSTM e considerando 3 semanas anteriores para a previsão, alcançando um RMSE de 37,038 e um MAPE de 2,434. É importante levar em conta, entretanto, que a falta de dados mais precisos dificultou alguns objetivos do trabalho. A falta de uma granularidade diária nos dados tornou a rede menos precisa, assim como a falta de algumas informações como o número total de leitos, que poderia servir de limite superior para as predições. Trabalhos futuros devem buscar fontes de dados mais ricas, para que estudos mais aprofundados, que levem em conta mais características do problema, possam ser feitos.

Referências

Academy, D. S. Deep learning book, https://www.deeplearningbook.com.br/arquitetura-de-redes-neurais-long-short-term-memory/, junho.

Beenhakker, H. L. (1963). Multiple correlation—a technique for prediction of future hospital bed needs. *Operations Research*, 11(5):824–839.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to time series and forecasting*. Springer.

Carvalho, D. M. T. (2009). Sistema de informações hospitalares do sus – sih-sus. *A experiência brasileira em sistemas de informação em saúde*, 49.

Chatfield, C. (2013). *The analysis of time series: theory and practice*. Springer.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonal-trend decomposition. *J. Off. Stat*, 6(1):3–73.

Coghlan, A. (2015). A little book of r for time series. *Published under Creative Commons Attribution*, 3.

Earnest, A., Chen, M. I., Ng, D., and Sin, L. Y. (2005). Using autoregressive integrated moving average (arima) models to predict and monitor the number of beds occupied during a sars outbreak in a tertiary hospital in singapore. *BMC Health Services Research*, 5(1):1–8.

Esling, P. and Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):1–34.

Gers, F. A., Schraudolph, N. N., and Schmidhuber, J. (2002). Learning precise timing with lstm recurrent networks. *Journal of machine learning research*, 3(Aug):115–143.

Graves, A. (2012). Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

IBGE. Projeção da população, https://cidades.ibge.gov.br/brasil/sc/panorama, junho.

IBGE. Sistema de informações hospitalares do sus – sih/sus, https://ces.ibge.gov.br/base-de-dados/metadados/ministerio-da-saude/sistema-de-informacoes-hospitalares-do-sus-sih-sus.html, abril.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kutafina, E., Bechtold, I., Kabino, K., and Jonas, S. M. (2019). Recursive neural networks in hospital bed occupancy forecasting. *BMC medical informatics and decision making*, 19(1):1–10.

Lessa, F. J. D., Mendes, A. d. C. G., Farias, S. F., Sá, D. A. d., Duarte, P. O., and Melo Filho, D. A. d. (2000). Novas metodologias para vigilância epidemiológica: uso do sistema de informações hospitalares-sih/sus. *Informe Epidemiológico do SUS*, 9:3–19.

Levcovitz, E. and Pereira, T. R. C. (1993). Sih/sus (sistema aih): uma análise do sistema público de remuneraçäo de internaçöes hospitalares no brasil-1983-1991. In *SIH/SUS (Sistema AIH): uma análise do sistema público de remuneraçäo de internaçöes hospitalares no Brasil-1983-1991*, pages 83–83.

Littig, S. J. and Isken, M. W. (2007). Short term hospital occupancy prediction. *Health care management science*, 10(1):47–66.

Robinson, A. and Fallside, F. (1987). *The utility driven dynamic error propagation network*. University of Cambridge Department of Engineering Cambridge.

Tello, M., Reich, E. S., Puckey, J., Maff, R., Garcia-Arce, A., Bhattacharya, B. S., and Feijoo, F. (2022). Machine learning based forecast for the prediction of inpatient bed demand. *BMC medical informatics and decision making*, 22(1):1–13.

Williams, R. J. and Zipser, D. (1995). Gradient-based learning algorithms for recurrent. *Backpropagation: Theory, architectures, and applications*, 433:17.