

Antônio Teixeira Santana Neto

**Modelagem estacionária do número de casos
e óbitos por Covid-19 no Brasil utilizando
perceptron multicamadas**

Viçosa, MG

2021

Antônio Teixeira Santana Neto

**Modelagem estacionária do número de casos e óbitos
por Covid-19 no Brasil utilizando perceptron
multicamadas**

Monografia apresentada ao Departamento de Engenharia Elétrica do Centro de Ciências Exatas e Tecnológicas da Universidade Federal de Viçosa, para a obtenção dos créditos da disciplina ELT 402 – Projeto de Engenharia II – e cumprimento do requisito parcial para obtenção do grau de Bacharel em Engenharia Elétrica.

Orientador Rodolpho Vilela Alves Neves

Viçosa, MG

2021

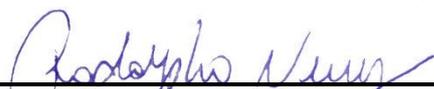
ANTÔNIO TEIXEIRA SANTANA NETO

**MODELAGEM ESTACIONÁRIA DO NÚMERO DE CASOS E ÓBITOS
PELA COVID-19 NO BRASIL UTILIZANDO O PERCEPTRON
MULTICAMADAS**

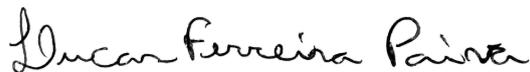
Monografia apresentada ao Departamento de Engenharia Elétrica do Centro de Ciências Exatas e Tecnológicas da Universidade Federal de Viçosa, para a obtenção dos créditos da disciplina ELT 402 – Projeto de Engenharia II – e cumprimento do requisito parcial para obtenção do grau de Bacharel em Engenharia Elétrica.

Aprovada em 27 de maio de 2021.

COMISSÃO EXAMINADORA



Prof. Dr. Rodolpho Vilela Alves Neves - Orientador
Universidade Federal de Viçosa



Eng. Dr. Lucas Ferreira Paiva - Coorientador
Universidade Federal de Viçosa



Prof. Dr. Leonardo Bonato Felix - Membro
Universidade Federal de Viçosa



Prof. Dr. Tiago Zanutelli - Membro
Instituto Federal do Espírito Santo

Este trabalho é dedicado a todos aqueles me apoiaram em minha trajetória.

Agradecimentos

Agradeço primeiramente à minha mãe, Cátia, e ao meu pai, Idelvan, por sempre terem incentivado os meus estudos e terem feito de tudo para que eu tivesse boas oportunidades. Obrigado por todo o apoio e incentivo durante todo esse tempo. Agradeço também ao meu irmão pela companhia presente em todo o caminho e, principalmente, durante o período em Viçosa. Agradeço à minha namorada, Raiane, por todo o apoio e parceria nos últimos anos, seu amor e afeto foram muito importantes pra mim e me deram forças para continuar em busca dos meus objetivos. Um agradecimento também à toda minha família por toda fé em mim.

Agradeço aos meus amigos da ELT, em especial Patrícia, Lucas (Nanuque), Matheus (Mathias), Leonardo (Belker), Thais e Adriana. Todas as conversas tomando café no Geraes e momentos que passamos juntos foram incríveis e, com certeza, minha trajetória pela UFV foi melhor com vocês.

Ao meu professor e orientador Rodolpho Vilela, por aceitar me orientar e, acima de tudo, pela paciência comigo e com meu trabalho, além do tempo dedicado.

Um agradecimento à equipe da Casa Amarela, em especial à Tia Dani, Denise e Mariza, e aos colegas tutores. Obrigado por todas as conversas, cafés, e por me proporcionarem um dos melhores lugares para se estar na universidade.

À Empresa Júnior diElétrica e todos os seus membros, por ter sido um ambiente de crescimento profissional e amadurecimento, além das amizades que conquistei.

Finalmente, agradeço aos professores e funcionários do departamento de Engenharia Elétrica, por todo o conhecimento transmitido ao longo da minha graduação.

“Não sei, só sei que foi assim”
(Chicó, O Auto da Compadecida)

Resumo

A pandemia do novo coronavírus tornou-se um dos grandes desafios do século XXI, trazendo grandes impactos socioeconômicos, além de levar os sistemas de saúde mundiais a um nível acima do limite. Dessa forma, ferramentas que possibilitem maior entendimento do comportamento da doença pode ser crucial durante a elaboração de estratégias e tomadas de decisões que visem a mitigar seus impactos. Nesse trabalho, redes neurais artificiais baseadas no Perceptron Multicamadas são desenvolvidas a fim de modelar o número máximo de pacientes infectados e número máximos de óbitos pela COVID-19 em oito grupos de estudo: a capital do estado de São Paulo; as capitais brasileiras somadas do Distrito Federal; as cinco macro regiões do Brasil; e, por fim, um grupo contendo todas as cidades e municípios brasileiros. O algoritmo de busca randomizada foi aplicado a fim de encontrar as melhores combinações de hiperparâmetros para cada objetivo em cada grupo de estudo. As redes neurais artificiais escolhidas são aquelas que apresentaram o maior coeficiente de determinação R^2 durante a etapa de validação cruzada. Os valores de R^2 das redes selecionadas durante o treinamento variam de 0,999 para os dois primeiros grupos de estudo até 0,32 para o último. Dentre todos os modelos, as arquiteturas selecionadas possuem o L-BFGS como otimizador de pesos sinápticos e a tangente hiperbólica como função de ativação de maior destaque.

Palavras-chaves: Rede perceptron multicamadas; Otimização de hiperparâmetros; COVID-19; SARS-CoV-2; Coronavírus.

Abstract

The pandemic of the new coronavirus has become one of the great challenges of the 21st century, bringing great socioeconomic impacts, besides pushing the world health systems to the limit. Thus, tools that allow a better understanding of the disease behavior can be crucial during the elaboration of strategies and decision-making aimed at mitigating its impacts. In this work, artificial neural networks based on multilayer Perceptron are developed to model the maximum number of infected patients and the maximum number of deaths by COVID-19 in eight study groups: the capital of the state of São Paulo; the Brazilian capitals plus the Federal District; the five macro-regions of Brazil; and, finally, a group containing all Brazilian cities and towns. The randomized search algorithm was applied in order to find the best combinations of hyperparameters for each objective in each group. The artificial neural networks chosen are those that presented the highest coefficient of determination R^2 during the cross-validation step. The R^2 values for the best networks range from 0.999 for the first two groups to 0.32 for the last. Among all the models, the best results generally have the LBFGS as the synaptic weights optimizer and the hyperbolic tangent as the activation function.

Key-words: Multilayer perceptron; Hyperparameter optimization; COVID-19; SARS-CoV-2; Coronavirus.

Lista de ilustrações

Figura 1 – Representação da rede <i>Perceptron</i> multicamadas.	18
Figura 2 – Ilustração das funções logística e tangente hiperbólica, com seus respectivos intervalos de saturação e variação dinâmica.	21
Figura 3 – Representação do processo de modelagem utilizando o PMC. Primeiramente, os dados são coletados e armazenados na forma de um <i>dataset</i> . Parte desses dados são usados para treinamento e validação de arquiteturas de PMC para cada combinação de hiperparâmetro, de modo a encontrar o melhor modelo possível. Os melhores modelos encontrados são usados para determinar os objetivos desejados.	23
Figura 4 – Curva do número de pacientes infectados na cidade de São Paulo. Comparação entre os dados reais e os modelados pelo PMC.	29
Figura 5 – Curva do número de pacientes que vieram a óbito na cidade de São Paulo. Comparação entre os dados reais e os modelados pelo PMC.	30
Figura 6 – Curva do número de pacientes infectados nas capitais brasileiras. Comparação entre os dados reais e os modelados pelo PMC.	31
Figura 7 – Curva do número de pacientes que vieram a óbito nas capitais brasileiras. Comparação entre os dados reais e os modelados pelo PMC.	32
Figura 8 – Curvas do número de pacientes infectados na região Norte do Brasil. Comparação entre os dados reais e os modelados pelo PMC.	34
Figura 9 – Curva do número de pacientes que vieram a óbito na região Norte do Brasil. Comparação entre os dados reais e os modelados pelo PMC.	34
Figura 10 – Curvas do número de pacientes infectados na região Centro Oeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.	36
Figura 11 – Curva do número de pacientes que vieram a óbito na região Centro Oeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.	36
Figura 12 – Curvas do número de pacientes infectados na região Sul do Brasil. Comparação entre os dados reais e os modelados pelo PMC.	38
Figura 13 – Curva do número de pacientes que vieram a óbito na região Sul do Brasil. Comparação entre os dados reais e os modelados pelo PMC.	38
Figura 14 – Curvas do número de pacientes infectados na região Sudeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.	40
Figura 15 – Curva do número de pacientes que vieram a óbito na região Sudeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.	41
Figura 16 – Curvas do número de pacientes infectados na região Nordeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.	42

Figura 17 – Curva do número de pacientes que vieram a óbito na região Nordeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.	43
Figura 18 – Curvas do número de pacientes infectados no Brasil. Comparação entre os dados reais e os modelados pelo PMC.	45
Figura 19 – Curva do número de pacientes que vieram a óbito no Brasil. Comparação entre os dados reais e os modelados pelo PMC.	45

Lista de tabelas

Tabela 1	– Quantidade de amostras de treino e teste em cada grupo de estudo. . .	24
Tabela 2	– Hiperparâmetros utilizados durante o treinamento das redes neurais artificiais. A primeira coluna representa o nome do hiperparâmetro, enquanto seus possíveis valores estão listados na segunda coluna. A última coluna representa as respectivas quantidades de parâmetros. . .	26
Tabela 3	– Resultados da etapa de validação para São Paulo/SP.	28
Tabela 4	– Hiperparâmetros para os melhores modelos referentes à São Paulo-SP. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.	29
Tabela 5	– Tempo de treinamento para cada objetivo, em minutos, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas. .	29
Tabela 6	– Resultados da etapa de validação para as capitais brasileiras.	30
Tabela 7	– Hiperparâmetros para os melhores modelos referentes às capitais brasileiras. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.	31
Tabela 8	– Tempo de treinamento para cada objetivo, em minutos, nas capitais brasileiras, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.	31
Tabela 9	– Resultados da etapa de validação para a região Norte do Brasil.	32
Tabela 10	– Tempo de treinamento para cada objetivo, em minutos, na região Norte, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.	33
Tabela 11	– Hiperparâmetros para os melhores modelos referentes à região Norte do Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.	33
Tabela 12	– Resultados da etapa de validação para a região Centro Oeste do Brasil.	34
Tabela 13	– Hiperparâmetros para os melhores modelos referentes à região Centro Oeste do Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.	35
Tabela 14	– Tempo de treinamento para cada objetivo, em minutos, na região Centro Oeste, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.	35

Tabela 15 – Resultados da etapa de validação para a região Sul do Brasil.	37
Tabela 16 – Hiperparâmetros para os melhores modelos referentes à região Sul do Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.	37
Tabela 17 – Tempo de treinamento para cada objetivo, em minutos, na região Sul, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.	39
Tabela 18 – Resultados da etapa de validação para a região Sudeste do Brasil. . . .	39
Tabela 19 – Hiperparâmetros para os melhores modelos referentes à região Sudeste do Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.	39
Tabela 20 – Tempo de treinamento para cada objetivo, em minutos, na região Sudeste, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.	40
Tabela 21 – Resultados da etapa de validação para a região Nordeste do Brasil. . . .	41
Tabela 22 – Hiperparâmetros para os melhores modelos referentes à região Nordeste do Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.	42
Tabela 23 – Tempo de treinamento para cada objetivo, em minutos, na região Nordeste, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.	43
Tabela 24 – Resultados da etapa de validação para o Brasil.	43
Tabela 25 – Hiperparâmetros para os melhores modelos referentes ao Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.	44
Tabela 26 – Tempo de treinamento para cada objetivo, em minutos, no Brasil, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.	44
Tabela 27 – Coeficiente de determinação R^2 de cada objetivo em cada grupo de estudo.	46

Lista de abreviaturas e siglas

CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
COVID-19	Coronavirus Disease 2019
DP	Desvio Padrão
EP	Erro Percentual
EPM	Erro Percentual Médio
EQM	Erro Quadrático Médio
IBGE	Instituto Brasileiro de Geografia e Estatística
L-BFGS	Limited-memory Broyden-Fletcher-Goldfarb-Shanno
MLP	Multilayer perceptron
OMS	Organização Mundial da Saúde
PMC	Perceptron Multicamadas
RNA	Rede Neural Artificial
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
UFV	Universidade Federal de Viçosa

Sumário

1	INTRODUÇÃO	13
1.1	Objetivo geral	16
1.2	Objetivos específicos	16
2	REFERENCIAL TEÓRICO	17
2.1	Score Padronizado	17
2.2	Rede Perceptron Multicamadas	17
2.3	Hiperparâmetros	19
2.3.1	Taxa de aprendizagem	19
2.3.2	Parâmetro de regularização	20
2.3.3	Tamanho das camadas escondidas	20
2.3.4	Função de ativação	20
2.3.5	Método de otimização dos pesos	20
2.4	Dados da COVID-19 no Brasil	22
3	MATERIAIS E MÉTODOS	23
3.1	Banco de Dados	23
3.2	Pré-processamento de Dados	24
3.3	Treinamento do Perceptron Multicamadas	24
3.4	Otimização de Hiperparâmetros	26
3.5	Avaliação do modelo	26
4	RESULTADOS E DISCUSSÃO	28
4.1	São Paulo/SP	28
4.2	Capitais brasileiras	29
4.3	Região Norte	32
4.4	Região Centro Oeste	33
4.5	Região Sul	36
4.6	Região Sudeste	38
4.7	Região Nordeste	41
4.8	Brasil	42
4.9	Análise Geral	45
5	CONSIDERAÇÕES FINAIS	47
	REFERÊNCIAS	48

1 Introdução

O coronavírus é um patógeno cujo alvo principal é o sistema respiratório humano. Surtos anteriores de coronavírus (CoV) incluem a síndrome respiratória aguda grave, denominada SARS-CoV, e a síndrome respiratória do Oriente Médio, conhecida como MERS-CoV, as quais foram caracterizadas como grandes agentes que ameaçam a saúde pública (ROTHAN; BYRAREDDY, 2020).

Em meados de Dezembro de 2019, casos de pneumonia de origem desconhecida foram identificados em Wuhan, capital da província de Hubei, na China. O patógeno foi identificado como sendo um novo betacoronavírus de RNA envelopado, o qual foi atualmente denominado síndrome respiratória aguda 2, ou SARS-CoV-2 (GUAN et al., 2020). Poucos meses depois a Organização Mundial da Saúde (OMS) declarou a doença do coronavírus (COVID-19) como uma emergência de saúde pública de interesse internacional (GUAN et al., 2020). Em 11 de Março de 2020, a OMS declara pandemia de COVID-19 com, até então, cerca de 148 mil casos confirmados da doença e pouco mais de 4500 óbitos ao redor do mundo, segundo Worldometers (2021).

Os sintomas da infecção pelo novo coronavírus aparecem após um tempo de incubação de, aproximadamente, 5,2 dias (LI et al., 2020). O período sintomático da doença até o óbito do paciente varia de 6 a 41 dias, com uma mediana de 14 dias, segundo Wang, Tang e Wei (2020). Entretanto, esse intervalo depende da idade e do estado do sistema imunológico do paciente, de modo que é maior para aqueles acima de 70 anos comparado aos menores que 70 anos (WANG; TANG; WEI, 2020). Os sintomas mais comuns reportados são febre, tosse, tosse produtiva (com muco ou catarro), fadiga ou mialgia, dispneia (falta de ar).

Os milhões de casos de SARS-CoV-2 ao redor do mundo geraram uma tensão sem precedentes nos sistemas de saúde, incluindo o aumento das taxas de admissão hospitalar e da demanda por leitos de unidades de terapia intensiva (UTI), suporte respiratório avançado e profissionais de saúde capacitados. O impacto da pandemia sobre o sistema de saúde de cada país tem sido diferente, a depender do equilíbrio entre a oferta e a demanda, o qual está associado à capacidade de expandir sua infraestrutura de acordo com o avanço da pandemia (RANZANI et al., 2021).

Recentemente, a OMS declarou que a América do Sul se tornou o novo epicentro da pandemia do novo coronavírus (FEUER, 2020), à medida que o Brasil se tornou um dos países mais afetado pela doença, ocupando o segundo lugar em número de casos confirmados com mais de 13 milhões de casos acumulado e pouco mais de 350 mil mortes confirmadas, até o dia 12 de Abril de 2021 (WORLDMETERS, 2021). O Brasil corresponde a um

país de área continental com cerca de 210 milhão de habitantes. As suas cinco macro-regiões (Norte, Centro-Oeste, Sul, Sudeste e Nordeste) são bastante heterogêneas entre si, principalmente no quesito socioeconômico, a qual é refletida na qualidade dos serviços de saúde regionais, incluindo a disponibilidade de leitos hospitalares e de profissionais treinados da área da saúde (RANZANI et al., 2021). Segundo Candido et al. (2020), a SARS-CoV-2 chegou ao país via voos internacionais, os casos ficaram, inicialmente, concentrados nas áreas metropolitanas, propagando das capitais para o interior do país.

De acordo com Massuda et al. (2018), as recentes crises econômicas e políticas intensificaram ainda mais os problemas estruturais do sistema de saúde unificado do Brasil (Sistema Único de Saúde) - cujo objetivo é proporcionar cobertura universal de saúde - incluindo lacunas na governança e organização, subfinanciamento crônico e baixa eficácia clínica. A pandemia da COVID-19 desafiou o sistema de saúde do país de tal modo que as disparidades regionais existentes provavelmente intensificaram devido à pandemia, afetando desproporcionalmente os grupos econômicos mais vulneráveis da população (RANZANI et al., 2021).

As últimas pandemias de SARS, em 2002 e 2003, foram controladas e, finalmente, contidas através da aplicação de medidas de controle, como restrições de viagens e isolamento de pacientes (CAR et al., 2020). Medidas de saúde pública que requerem distanciamento social, controle comunitário e fechamento de escolas e empresas foram implementadas no Brasil e em outros países do mundo, a fim de diminuir a transmissão do vírus.

Além disso, algumas vacinas já foram produzidas e estão sendo disponibilizadas mundialmente, mesmo que de forma desigual. No Brasil estão presentes vacinas de farmacêuticas como a AstraZeneca (KNOLL; WONODI, 2021), produzida em parceria com a Fundação Oswaldo Cruz (Fiocruz), da Pfizer/BioNTech (BRITTON et al., 2021) e da farmacêutica Sinovac (WU et al., 2021), a qual está sendo produzida em parceria com o Instituto Butantan. O país já conta com mais de 30 milhões de vacinados até o dia 12 de Abril de 2021 (COTA, 2020). Todavia, devido à vacinação lenta, o número de casos confirmados e óbitos continua a subir no território nacional.

Nesse contexto, ter em mãos métodos confiáveis de predição e modelagem da propagação da COVID-19 seria de grande benefício para convencer a opinião pública da importância de aderir e respeitar as medidas de contenção da doença (CAR et al., 2020), além de possibilitar o planejamento e ajudar na tomada de decisões no que diz respeito ao sistema de saúde pública (SAVI; SAVI; BORGES, 2020).

Apesar de recente, a literatura atual conta com vários projetos que buscam modelar a propagação do novo coronavírus ou analisar as dinâmicas da doença. Savi, Savi e Borges (2020) propõem um modelo matemático epidemiológico baseado no *Epidemiological Model Susceptible, Exposed, Infected and Recovered* (SEIR), bastante utilizado para análise e

compreensão de epidemias. A proposta faz uso dos dados disponíveis sobre a doença em países como China, Itália, Irã e Brasil, e posteriormente analisa diversos cenários da pandemia no Brasil, reforçando o fato de que medidas governamentais de controle somadas a atitudes individuais de proteção são essenciais para reduzir o número de infectados pela doença e, assim, o tempo de pandemia. Além disso, os resultados mostraram que o fim precipitado das medidas de contenção e isolamento social podem contribuir para o aumento significativo do número de infectados na população. [Zhao et al. \(2020\)](#) propôs análises estatísticas baseadas no conceito de Poisson a fim de estimar o número real de casos de COVID-19 que não haviam sido reportados na primeira quinzena de Janeiro de 2020. Segundo eles, houve cerca de 469 casos não relatados de 1 a 15 de Janeiro de 2020, vindo a aumentar cerca de 21 vezes após 17 de Janeiro.

Inteligência Artificial (IA) é o estudo e desenvolvimento de ferramentas que imitam a inteligência humana. Essa técnica tem tido uma grande sucesso em uma gama de campos de estudo como detecção de fraudes, visão computadorizada, publicidade online, robótica, *drivers* automáticos, entre outros. Com seu sucesso em áreas como diagnóstico de doenças, tratamento e monitoramento de pacientes, descoberta de medicamentos e epidemiologia, etc, há uma grande esperança que a IA possa ser uma ótima área de pesquisa para enfrentar os desafios que o ser humano enfrenta atualmente ([NAJARAN, 2020](#)). São diversos os campos de aplicação de IA no que se refere ao SARS-CoV-2, mas os principais englobam aplicações clínicas, processamento de imagens relacionadas à COVID-19 como raios-X e tomografia computadorizada, além de estudos epidemiológicos ([NAJARAN, 2020](#)).

As redes neurais artificiais (RNA) tem sido aplicadas em diversas áreas de estudo como ciências ambientais, agricultura, finanças, epidemiologia e saúde pública ([MOLLALO; RIVERA; VAHEDI, 2020](#)). Alguns dos principais pontos a favor das RNA são sua capacidade de se adaptar aos dados mesmo com pouco conhecimento sobre o comportamento dos mesmos, além da sua grande capacidade de generalização, atuando como um excelente aproximador de funções ([RIZK-ALLAH; HASSANIEN, 2020](#); [IBRAHIM, 2020](#)).

Enquanto abordagens estatísticas tradicionais podem oferecer modelos precisos, técnicas de IA como as RNA podem ser a chave para encontrar modelos preditivos de alta qualidade em se tratando da pandemia do novo coronavírus ([CAR et al., 2020](#)).

Nesse contexto, o presente trabalho propõe modelar, separadamente, o número máximo de casos confirmados e mortes confirmadas pela COVID-19 no Brasil, a partir da aplicação de redes Perceptron Multicamadas. O estudo faz uso de dados de uma série temporal com dados diários sobre os casos de COVID-19 (número de pacientes infectados e de mortes) no Brasil por localidade. Os dados são então transformados em entradas e saídas para serem utilizados no processo de treinamento das RNAs. O algoritmo de busca randomizada é aplicado a fim de encontrar as melhores combinações de hiperparâmetros que produzam os melhores resultados com base no coeficiente de determinação R^2 . Esse

estudo pode apresentar resultados promissores que auxiliem a tomada de decisões no que diz respeito ao enfrentamento da doença no país.

1.1 Objetivo geral

Este trabalho tem como objetivo modelar, separadamente, o número máximo de casos confirmados e mortes confirmadas pela COVID-19 no Brasil, a partir da aplicação de redes Perceptron Multicamadas com o uso de dados informados pelo projeto de pós-doutorado "Monitoramento contínuo da COVID-19 no Brasil: coleta, análise e modelagem de dados epidêmicos".

1.2 Objetivos específicos

Os objetivos específicos do projeto são:

- Selecionar um banco de dados que contenha as informações sobre número de pacientes por localização no Brasil;
- Definir as respectivas entradas e saídas da rede neural. Separar os dados entre grupos de estudo para aplicação dos algoritmos. Efetuar o pré-processamento das entradas e saídas;
- Selecionar a melhor rede neural artificial (RNA) para estimar o número máximo de pacientes em cada grupo com base no valor do coeficiente de determinação R^2 durante a etapa de treinamento.

2 Referencial Teórico

Este capítulo contém a fundamentação teórica dos métodos adotados no modelo de estimação do número de pacientes (quantidade de infectados e de óbitos) da COVID-19 no Brasil. Será feita uma breve abordagem sobre a ferramenta de normalização adotada, Teste-Z, seguida da conceituação de redes neurais artificiais e a apresentação de uma das primeiras redes, a rede *Perceptron*. Posteriormente será tratado do *Perceptron* multicamadas, a qual foi a rede neural adotada neste trabalho. Por fim, tem-se a definição de hiperparâmetros e aqueles que serão utilizados neste estudo.

2.1 Score Padronizado

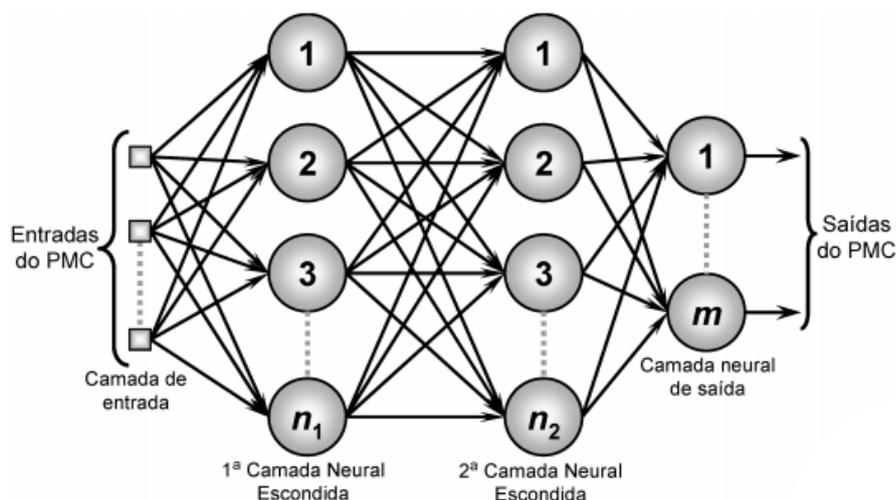
Score padronizado, também conhecido como Teste-Z (ou *Z-Score*), é uma das técnicas de normalização de dados mais utilizada, principalmente em se tratando de modelos de redes neurais artificiais (JAIN; SHUKLA; WADHVANI, 2018). Além disso, é um método bastante útil quando os valores mínimos e máximos dos dados de estudo não são conhecidos (MOHAMAD; USMAN, 2013).

Essa técnica, utiliza a média e o desvio padrão de um conjunto de dados para normalizar cada amostra que servirá de entrada para a rede neural artificial. A transformação pode ser dada por (JAIN; SHUKLA; WADHVANI, 2018), onde μ é a média aritmética e σ é o desvio padrão do conjunto de dados. Como resultado, cada conjunto de dados passará a ter média zero e desvio padrão unitário (DZIERŻAK, 2019).

$$x'_i = \frac{x_i - \mu}{\sigma} \quad (2.1)$$

2.2 Rede Perceptron Multicamadas

O Perceptron Multicamadas consiste de um sistema de múltiplos neurônios, ou nós, interconectados, caracterizado pela presença de pelo menos uma camada intermediária (escondida) de neurônios, situada entre a camada de entrada e a respectiva camada neural de saída, conforme representado na Figura 1. As conexões entre os neurônios artificiais são ponderados por pesos sinápticos e sinais de saída, os quais são modificados por meio de funções de ativação, o que permite o sistema aproximar problemas complexos e não lineares (GARDNER; DORLING, 1998). Além disso, nesse tipo de rede, a informação segue apenas uma direção, da camada de entrada até a camada de saída, caracterizando uma arquitetura *feed-forward*.

Figura 1 – Representação da rede *Perceptron* multicamadas.

Fonte: [Silva, Spatti e Flauzino \(2016\)](#).

O processo de treinamento, ajuste dos pesos entre as conexões de modo a minimizar uma função de erro (geralmente o erro quadrático médio) ([HIPPERT; PEDREIRA; SOUZA, 2001](#)), de um perceptron multicamadas se dá por meio do algoritmo conhecido como *backpropagation* ([RUMELHART; HINTON; WILLIAMS, 1986](#)). Essa técnica utiliza do procedimento do gradiente descendente para localizar o ponto de mínimo local, ou global, da superfície da função de erro. Inicialmente, os pesos da rede são inicializados aleatoriamente com valores pequenos, de modo a selecionar um ponto qualquer da superfície de erro. O algoritmo, então, calcula o gradiente local e atualiza os pesos de modo a deslocar em direção ao mínimo global. Dada uma superfície de erro contínua, é esperado que os pesos irão convergir de modo a atingir o mínimo local da superfície de erro ([GARDNER; DORLING, 1998](#)).

O algoritmo do *backpropagation* é resumido abaixo e os detalhes de sua implementação podem ser encontrado na maioria dos livros ([Bishop \(1995\)](#); [Silva, Spatti e Flauzino \(2016\)](#)) sobre redes neurais artificiais.

Algoritmo de treinamento do PMC:

1. Inicializar os pesos da rede neural,
2. Alimentar a rede com a primeira amostra de entrada dos dados de treino,
3. Propagar o vetor de entrada através da rede para obter a saída,
4. Calcular o sinal de erro por meio da comparação do sinal de saída com o respectivo valor desejado,
5. Efetuar a propagação reversa do sinal de rede ao longo da rede,

6. Ajustar os pesos a fim de minimizar o erro geral,
7. Repetir os passos 2 ao 6 com a próxima amostra de entrada, de modo que o erro geral seja satisfatoriamente pequeno.

Finalmente, após o processo de treinamento, a rede neural pode ser validada com sua aplicação nos dados de teste, os quais consistem de dados que não foram utilizados no processo de treino, portanto, desconhecidos pela rede.

O perceptron multicamadas é capaz, dessa maneira, de aproximar funções não-lineares sem que seja preciso ter conhecimento prévio da natureza da relação entre entradas e saídas. Essa é uma das vantagens do PMC sobre as regressões convencionais. Se a relação entre entrada e saída é não-linear, então, a regressão linear se torna uma ferramenta inapropriada para a aplicação em questão (GARDNER; DORLING, 1998).

2.3 Hiperparâmetros

Existe uma grande variedade de hiperparâmetros (do inglês, *hyperparameters*), mas todos compartilham o fato de que não há um método geral de aproximação que seja capaz de otimizá-los. Desse modo, esse ajuste se dá sobre valores pré-estabelecidos (Schilling et al., 2015).

Treinar e testar um modelo de rede neural artificial (RNA) requer determinar a estrutura da rede (número e tamanho das camadas escondidas) além de parâmetros como a taxa de aprendizado e taxa de decaimento. Esses tipos de parâmetros são chamados de hiperparâmetros, visto que eles devem ser determinados antes da etapa de treinamento da rede (Diaz et al., 2017). Em projetos de aprendizado de máquina, ajustar esses hiperparâmetros é a chave para reduzir tempo computacional, gerando um erro considerável (KARAKI; IVANOV, 2020). Assim sendo, a seguir tem-se uma descrição mais extensa sobre os hiperparâmetros utilizados nesse trabalho.

2.3.1 Taxa de aprendizagem

A taxa de aprendizagem η exprime quão rápido o processo de treinamento da rede estará sendo conduzido rumo à sua convergência (estabilização) (SILVA; SPATTI; FLAUZINO, 2016). Este, na maioria das vezes, é o hiperparâmetro mais importante e sua escolha deve ser realizada com cautela a fim de evitar instabilidade no processo de treinamento. Valores típicos estão no intervalo $10^{-6} < \eta < 1$, mas aqueles mais apropriados vai depender da aplicação. O valor padrão de 0,01 geralmente produz resultados aceitáveis em projetos padrões de perceptron multicamadas (BENGIO, 2012).

2.3.2 Parâmetro de regularização

Há situações em que o aumento do número de neurônios, assim como o incremento de camadas intermediárias podem levar a saída do PMC para a circunstância de memorização excessiva, na qual este acaba decorando suas respostas frente aos estímulos introduzidos em suas entradas. Nessas ocorrências, o erro durante a fase de aprendizado tende a ser bem baixo; contudo, durante a fase de generalização frente aos subconjuntos de teste, o erro tende a assumir valores bem elevados, fato que denota a condição de *overfitting* (SILVA; SPATTI; FLAUZINO, 2016).

O parâmetro de regularização $L2$ limita a influência das amostras de entrada, a fim de evitar que a RNA seja treinada com algum viés advindo de uma amostra que tenha muita correlação com a saída. Quanto maior o parâmetro, menor será essa influência (CAR et al., 2020). Ou seja, é um hiperparâmetro que contribui para reduzir o *overfitting* (BENGIO, 2012).

2.3.3 Tamanho das camadas escondidas

Corresponde à quantidade de camadas escondidas e as respectivas quantidades de neurônios em cada camada. O uso de camadas escondidas com a mesma quantidade de neurônios em cada uma delas pode trazer resultados melhores, ou iguais, quando comparados a redes que possuem suas camadas escondidas em formato de pirâmide, isto é, com número crescente (ou decrescente) de neurônios em cada camada. Contudo, é importante frisar que essa característica pode variar conforme a aplicação (LAROCHELLE et al., 2009). Neste trabalho, serão empregadas arquiteturas de PMC com uma única camada escondida, contendo de 1 a 40 neurônios.

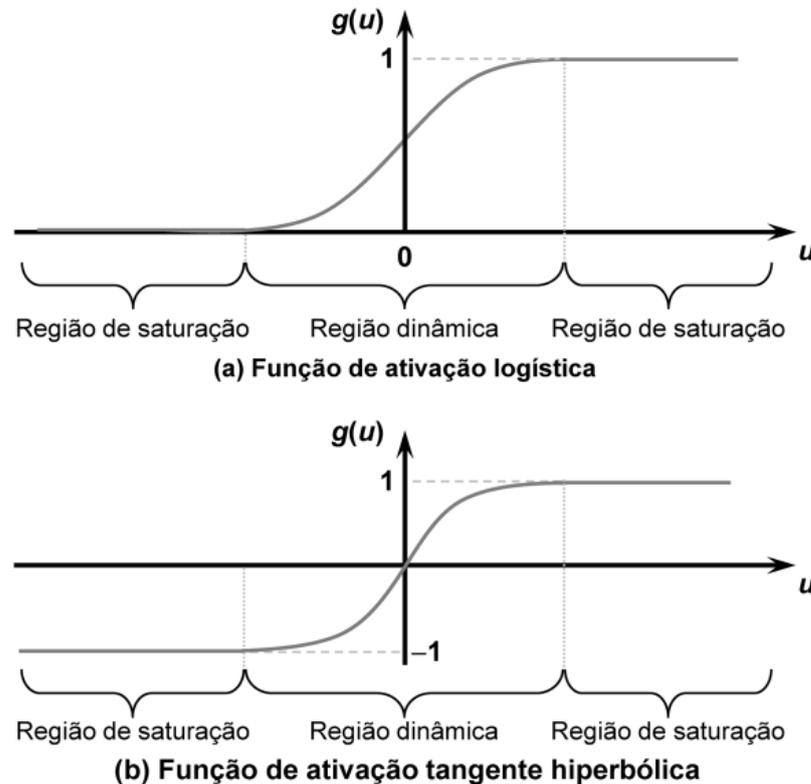
2.3.4 Função de ativação

Representa a função aplicada à saída de cada neurônio da rede neural artificial. São funções diferenciáveis e que trazem a não-linearidade ao modelo. As funções mais utilizadas são a logística e tangente hiperbólica (MAIER; DANDY, 2000), representadas pela Figura 2. É importante notar que técnicas que normalizam as entradas da rede, considerando os limites numéricos produzidos pelas funções de ativação adotadas, visam a melhorar o desempenho computacional do processo de treinamento (SILVA; SPATTI; FLAUZINO, 2016).

2.3.5 Método de otimização dos pesos

Corresponde ao algoritmo utilizado para recalculer os pesos do PMC durante o processo de *backpropagation* na fase de treinamento (CAR et al., 2020). Neste trabalho, foi empregado dois métodos para otimização dos pesos: *lbfgs* e *adam*.

Figura 2 – Ilustração das funções logística e tangente hiperbólica, com seus respectivos intervalos de saturação e variação dinâmica.



Fonte: [Silva, Spatti e Flauzino \(2016\)](#).

O primeiro corresponde ao algoritmo L-BFGS, o qual é uma variação do método newtoniano para minimização de funções ([WU et al., 2018](#)). Dentre todos aqueles que derivam do método newtoniano, este é o que apresenta melhor performance com menor uso de memória ([Popa, 2015](#)), o que implica em maior velocidade de processamento ([WU et al., 2018](#)). Para aplicações com banco de dados menores, esse método pode convergir mais rápido com melhor performance ([PEDREGOSA et al., 2011](#); [BUTINCK et al., 2013](#)).

Por outro lado, o *Adam* corresponde a um método de otimização de pesos baseado na estratégia do gradiente descendente. Todavia, diferentemente do método gradiente descendente estocástico, o primeiro requer apenas a derivada de primeira ordem do gradiente. Assim, ele calcula taxas de aprendizado adaptativas individuais para diferentes parâmetros a partir de estimativas do primeiro e segundo momentos dos gradientes ([KINGMA; BA, 2015](#)). Algumas de suas vantagens são sua eficiência computacional, pouco uso de memória além de adaptar-se bem para problemas com enorme quantidade de dados e parâmetros ([KINGMA; BA, 2015](#)).

2.4 Dados da COVID-19 no Brasil

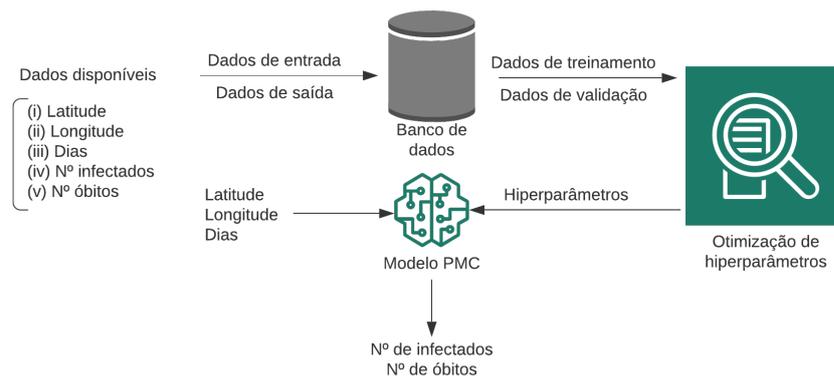
Desde a explosão da pandemia no mundo, a COVID-19 tem sido constantemente monitorada por vários países e organizações. A base de dados utilizada nesse estudo é parte do projeto de pós-doutorado Monitoramento contínuo da COVID-19 no Brasil: coleta, análise e modelagem de dados epidêmicos, registrado na Pró-Reitoria de Pesquisa e Pós-Graduação da Universidade Federal de Viçosa, com bolsa da CAPES (COTA, 2020). Os dados consistem de números relacionados à COVID-19 no Brasil. Há dados de casos e óbitos por município, com informações oficiais do Ministério da Saúde, juntamente com os das Secretarias Estaduais de Saúde obtidos pelo Brasil.IO, coletados entre 25 de Fevereiro de 2020 até 12 de Abril de 2021 de 5570 municípios e cidades brasileiras.

Os dados foram publicados como uma série temporal e mostra a propagação da doença em vários locais ao longo do tempo. A fim de treinar o PMC, os dados foram rearranjados para serem as entradas e saída da rede. Para cada caso registrado, as respectivas latitude e longitude da localização, assim como a data no qual foi contabilizado foram tomadas. A data foi convertida em número de dias desde o primeiro caso reportado. Dessa forma, cada amostra do *dataset* contém informações sobre o número de pacientes (infectados e pacientes mortos) em uma dada localização em um determinado dia desde o primeiro caso reportado. Longitude, latitude e o número de dias desde o primeiro caso foram usados como entradas da rede neural artificial, enquanto a saída representa a quantidade total de pacientes em cada grupo (casos confirmados e mortes confirmadas).

3 Materiais e Métodos

Neste capítulo são apresentados os materiais utilizados e os métodos adotados para modelar o número de casos confirmados de infecção por COVID-19 e o número de mortes confirmadas. Todas as etapas desde a seleção dos dados, pré-processamento, modelagem dos modelos e, finalmente, ferramentas de validação dos resultados são descritas a seguir. Uma visão geral do processo de todo o processo de modelagem é dado na Figura 3.

Figura 3 – Representação do processo de modelagem utilizando o PMC. Primeiramente, os dados são coletados e armazenados na forma de um *dataset*. Parte desses dados são usados para treinamento e validação de arquiteturas de PMC para cada combinação de hiperparâmetro, de modo a encontrar o melhor modelo possível. Os melhores modelos encontrados são usados para determinar os objetivos desejados.



Fonte: Elaborado pelo autor.

3.1 Banco de Dados

Para este trabalho, os dados sobre a COVID-19 no Brasil foram divididos em oito grupos de estudo a fim de avaliar a capacidade de generalização da RNA: o primeiro grupo compreende os dados da cidade de São Paulo-SP, visto que foi o local no qual reportou-se o primeiro caso; o segundo compreende os dados de todas as capitais brasileiras mais o Distrito Federal, uma vez que representam os locais de maior concentração populacional; o terceiro ao sétimo grupo refere-se às cinco regiões geográficas do Brasil; por fim, os dados referentes a todas as cidades e municípios brasileiros. Em cada grupo, os dados foram aleatoriamente divididos em 5 partições (*folds*) iguais. E em cada uma dessas partições as amostras foram divididas entre dados de treino e dados de teste utilizando a proporção de 80%/20%, respectivamente. Detalhes sobre cada grupo de estudo estão na Tabela 1.

Tabela 1 – Quantidade de amostras de treino e teste em cada grupo de estudo.

Grupo de estudo	Amostras de treino	Amostras de teste
São Paulo/SP	330	83
Capitais brasileiras	8920	2231
Região Norte	142560	35640
Região Centro Oeste	150187	37547
Região Sul	380167	95042
Região Sudeste	467040	116760
Região Nordeste	578385	144597
Brasil	1840328	460082

3.2 Pré-processamento de Dados

Antes mesmo de iniciar o processo de treinamento da rede neural com as amostras de entrada e respectivas saídas desejadas é preciso aplicar ferramentas de pré-processamento de dados a fim de obter uma maior performance do modelo. O processo de normalização dos dados é uma dessas ferramentas que possuem grande impacto na eficiência da rede neural artificial (JAYALAKSHMI; A., 2011). Tal etapa consiste em "equalizar" os dados, deixando-os dentro de um intervalo que vai depender do método de normalização aplicado. Com isso, observa-se uma redução do número de interações na etapa de treino e do erro final, assim como diminuição do tempo computacional (Sola; Sevilla, 1997).

Com isso em vista, para este trabalho o método de normalização adotado foi o Score Padronizado, ou Teste-Z. Essa estratégia utiliza a média e o desvio padrão de cada variável para normalizar os valores de cada amostra individual e foi implementada através do uso da função *StandardScaler* presente no módulo *preprocessing* do *scikit-learn* (PEDREGOSA et al., 2011).

3.3 Treinamento do Perceptron Multicamadas

Para este trabalho, empregou-se o *Perceptron* Multicamadas a fim de modelar a propagação da COVID-19 no Brasil, retornando os números de pacientes infectados e aqueles que vieram a óbito. A escolha do PMC se deve ao fato de sua fácil implementação, além de promover modelos de alta qualidade mantendo o tempo de treinamento relativamente baixo (CAR et al., 2020). Além disso, essa arquitetura de rede possui como característica a sua enorme capacidade de aproximar funções, principalmente aquelas com relações não-lineares (SILVA; SPATTI; FLAUZINO, 2016).

Neste trabalho, os modelos de PMC que serão treinados possuem 3 neurônios de entrada (latitude, longitude e dias desde o primeiro caso reportado). A camada de saída consiste de um único neurônio, o qual fornecerá o número de pacientes para cada modelo

(número de casos confirmados e número de óbitos). O número de camadas escondidas com as respectivas quantidades de neurônios serão dados pelo algoritmo de otimização de hiperparâmetros.

A fim de confirmar a validade dos resultados, empregou-se o processo de validação cruzada, denominado k -partições (*k-fold cross-validation*). Essa é uma das técnicas estatísticas mais comumente utilizadas para seleção das melhores topologias candidatas (VABALAS et al., 2019), cujo propósito é avaliar a aptidão de cada uma quando aplicadas a um conjunto de dados que seja diferente daquele usado no ajuste de seus parâmetros internos (SILVA; SPATTI; FLAUZINO, 2016).

No método empregado, realiza-se a divisão do conjunto total de amostras em k -partições, sendo que $(k-1)$ delas serão usadas para compor o subconjunto de treinamento, ao passo que a partição restante constituirá o subconjunto de teste (SILVA; SPATTI; FLAUZINO, 2016). Por conseguinte, o processo de aprendizado se repete k vezes até que todas as partições tenham sido utilizadas como subconjunto de teste (SILVA; SPATTI; FLAUZINO, 2016). Neste trabalho, o valor empregado para o parâmetro k é igual a 5. O desempenho global de cada topologia candidata será obtido em função da média entre os desempenhos individuais observados quando da aplicação das k partições.

Além desse método, implementou-se, também, o procedimento de parada antecipada ou prematura (*early stopping*). Aqui, o processo de aprendizagem para uma topologia candidata é constantemente checado pela aplicação dos subconjuntos de teste (correspondendo a 10% das amostras de treinamento), sendo finalizado quando começar a haver elevação do erro (frente aos subconjuntos de teste) entre épocas sucessivas (SILVA; SPATTI; FLAUZINO, 2016).

Ademais, como várias topologias são testadas no mesmo conjunto de dados, torna-se, então, possível reutilizar aspectos dos modelos anteriores para iniciar os modelos posteriores. Esse método é chamado de *warm-start* (PEDREGOSA et al., 2011; BUITINCK et al., 2013). Neste trabalho, empregou-se essa prática em todos os grupos de estudo e para cada objetivo, de modo a reduzir o tempo computacional durante o processo de treinamento e otimização de hiperparâmetros.

A solução proposta foi implementada na linguagem de programação Python 3.8, com o uso da API *scikit-learn* (PEDREGOSA et al., 2011). Essa biblioteca foi utilizada pois permite a implementação de ferramentas de *machine learning* com facilidade, além de possuir a maioria dos métodos utilizados nesse trabalho. O treinamento das redes foi efetuado utilizando o *Compute Engine* da Google Cloud, no qual pode-se executar máquinas virtuais, as quais correspondem a *softwares* onde é possível executar programas e sistemas operacionais, armazenar dados, conectar-se a redes e executar outras funções de computação. Foi utilizada uma MV do tipo N1 com um processador escalonável Intel Xeon com 16 CPUs virtuais (vCPU) e 60Gb de memória RAM.

3.4 Otimização de Hiperparâmetros

Conforme dito anteriormente, os hiperparâmetros são valores que definem a arquitetura de uma rede neural artificial. A escolha de valores adequados é crucial para se obter modelos mais precisos e confiáveis. A determinação das melhores combinações pode ser feita de diversas formas, algumas delas são a estimação manual dos valores, a aplicação do algoritmo de busca exaustiva, ou por meio do algoritmo de busca randomizada.

Apesar da busca exaustiva e manual serem bastante utilizadas, a busca randomizada apresenta resultados mais eficientes, sendo capaz de encontrar modelos melhores, ou iguais, em comparação às outras duas metodologias utilizando de um menor tempo computacional (BERGSTRA; BENGIO, 2012).

Neste contexto, empregou-se a estratégia de busca randomizada (*Random Search*) para determinação das melhores combinações de hiperparâmetros. Ao todo foram testadas 30 combinações. Os possíveis valores desses parâmetros são dados na Tabela 2.

Tabela 2 – Hiperparâmetros utilizados durante o treinamento das redes neurais artificiais. A primeira coluna representa o nome do hiperparâmetro, enquanto seus possíveis valores estão listados na segunda coluna. A última coluna representa as respectivas quantidades de parâmetros.

Hiperparâmetro	Possíveis valores	Quantidade
Otimizador	Adam, LBFGS	2
Taxa de aprendizagem	0,1, 0,01, 0,5, 0,001, 0,0001	5
Parâmetro de regularização	0,01, 0,1, 0,001, 0,5, 0,0001	5
Tamanho das camadas escondidas	(1), (2), (3), (4), ..., (40)	40
Função de ativação	ReLU, identity, logistic, tanh	4
Combinações de hiperparâmetros		8000

3.5 Avaliação do modelo

A fim de avaliar o desempenho de cada modelo deste trabalho, empregou-se uma ferramenta estatística simples, porém bastante utilizada na análise de modelos de regressão: o coeficiente de determinação R^2 . Pode ser definida como o quão bem a variância dos dados observados pode ser explicada pelos dados previstos pelo modelo proposto (NAGELKERKE, 1991). Os valores de saída da rede neural (o número real de pacientes infectados ou que vieram a óbito) são contidos no vetor y , enquanto aqueles previstos pelo modelo são definidos pelo vetor \hat{y} . Assim sendo, o coeficiente de determinação R^2 pode ser determinado

como o coeficiente entre a variância residual e a variância total (CAR et al., 2020):

$$R^2 = 1 - \frac{S_{residual}}{S_{total}} = 1 - \frac{\sum_{i=0}^m (y_i - \hat{y}_i)^2}{\sum_{i=0}^m (y_i - 1/m \sum_{i=0}^m y_i)^2}, \quad (3.1)$$

com m representando o número total de amostras (tamanho dos vetores y e \hat{y}). Os valores para esse coeficiente estão no intervalo $0 \leq R^2 \leq 1$, onde um valor igual a zero significa que os dados previstos pelo modelo não conseguem explicar a variância dos dados reais, e um valor de R^2 igual a 1.0 representa o melhor resultado, de modo que os dados previstos conseguem explicar toda a variância dos dados reais (MENARD, 2000).

Devido ao uso da ferramenta de validação cruzada, cada arquitetura foi treinada cinco vezes, em bancos de dados diferentes. Assim, os resultados da validação cruzada são dados pelas médias dos valores de R^2 calculados. Os desvios padrão entre os diferentes valores de R^2 em cada partição também foram calculados.

4 Resultados e Discussão

Neste capítulo é apresentado os resultados obtidos na etapa de teste para cada um dos oito grupos de estudos, com base na metodologia descrita nas seções anteriores. Após apresentação, os resultados serão discutidos.

4.1 São Paulo/SP

Nesta seção são apresentados os resultados referentes ao banco de dados da capital de São Paulo. Vale lembrar que aqui foi relatado o primeiro caso de COVID-19 no país. Além disso, a capital possui a maior concentração populacional do país juntamente com o maior número de casos confirmados da doença, 663212 pacientes infectados e 24282 óbitos, durante os primeiros 412 dias de pandemia no país.

Para cada objetivo (número de pacientes infectados e número de óbitos) foram treinadas 150 redes neurais. O melhor modelo apresentou um coeficiente de determinação R^2 médio entre as partições de 0,9997 durante a etapa de treinamento, para ambos os objetivos. Porém, a fim de validar as arquiteturas é necessário aplicá-las aos dados de teste, os quais correspondem a novas amostras, desconhecidas pelas redes. Os resultados referentes a essa etapa de teste estão presentes na Tabela 3, na qual é possível observar os respectivos valores de R^2 , erro médio absoluto (*MAE*) e a raiz quadrada do erro-médio (*RMSE*).

Tabela 3 – Resultados da etapa de validação para São Paulo/SP.

Objetivo	R^2	MAE	RMSE
Infectados	0,9999	0,0074	0,0099
Óbitos	0,9989	0,0257	0,0321

Foram treinadas 30 diferentes combinações de hiperparâmetros para cada objetivo. Os valores escolhidos para as arquiteturas são representados na Tabela 4.

Observa-se na Tabela 4 que, para ambos os modelos, o LBFGS se destacou como melhor opção para otimização de pesos sinápticos. Todavia, os outros hiperparâmetros diferente entre si quanto aos valores ótimos para cada objetivo, de modo para o número de infectados e número de óbitos, as funções de ativação escolhidas foram ReLU e tanh (tangente hiperbólica), respectivamente. O tempo de treinamento para cada caso está representado na Tabela 5.

Com base nesses valores, é possível estimar a curva de comportamento da COVID-19 em cada caso a partir dos modelos desenvolvidos. As Figuras 4 e 5 representam essa

Tabela 4 – Hiperparâmetros para os melhores modelos referentes à São Paulo-SP. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.

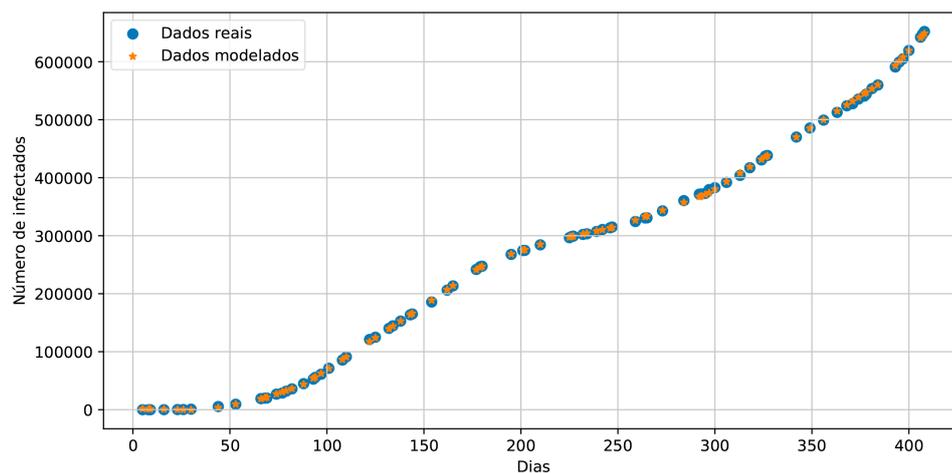
Hiperparâmetro	Modelo para n ^o de infectados	Modelo para n ^o de óbitos
Otimizador	LBFGS	LBFGS
Taxa de aprendizagem	0,001	0,01
Parâmetro de regularização	0,001	0,0001
Tamanho das camadas escondidas	37	17
Função de ativação	ReLU	tanh

Tabela 5 – Tempo de treinamento para cada objetivo, em minutos, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.

Objetivo	Tempo de treinamento (min)
Número de infectados	0,3471
Número de óbitos	0,2424
Média	0,2947

comparação entre os dados reais e aqueles modelados pelas arquiteturas de PMC. Assim sendo, constata-se a capacidade dos modelos em retratar, com erro mínimo, cada objetivo.

Figura 4 – Curva do número de pacientes infectados na cidade de São Paulo. Comparação entre os dados reais e os modelados pelo PMC.

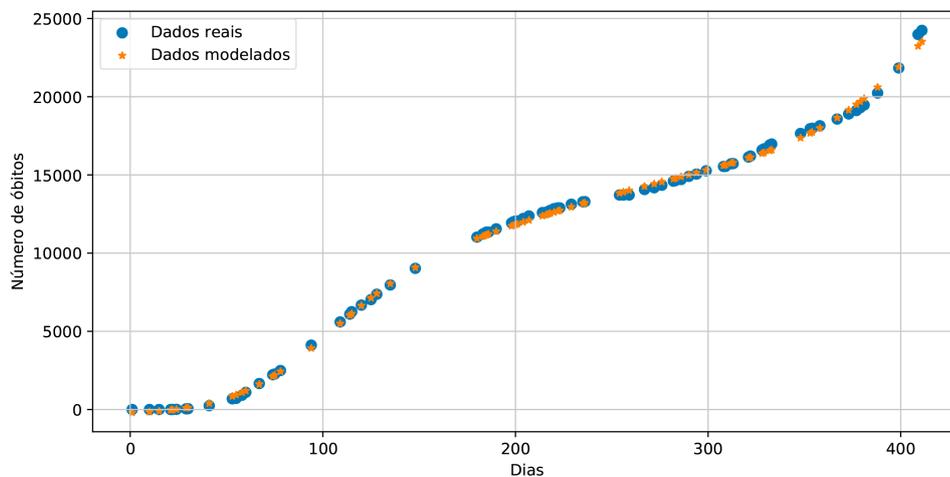


Fonte: Elaborado pelo autor.

4.2 Capitais brasileiras

Nesta seção são apresentados os resultados do grupo de estudo referentes às 26 unidades federativas do Brasil, somado do Distrito Federal. Essas cidades representam

Figura 5 – Curva do número de pacientes que vieram a óbito na cidade de São Paulo. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

regiões de grande concentração populacional, com alta circulação de pessoas. Durante a etapa de treinamento, as redes neurais foram alimentadas com 8920 amostras. Já na etapa de teste, os modelos formados pela melhor combinação de hiperparâmetros foram validados com 2231 amostras novas, desconhecidas pelas redes neurais.

Para cada objetivo o algoritmo de busca randomizada efetuou 30 interações, para cada interação foi aplicado o processo de validação cruzada com 5 partições, totalizando, assim, 150 redes neurais treinadas. O melhor modelo apresentou um coeficiente de determinação R^2 médio de 0,9975 e 0,9979 para o número de pacientes infectados e óbitos, respectivamente. As arquiteturas escolhidas foram validadas com os dados de teste. Os resultados dessa etapa estão presentes na Tabela 6, na qual pode-se observar os valores de R^2 , erro médio absoluto (MAE) e a raiz quadrada do erro-médio ($RMSE$). Os valores obtidos dos hiperparâmetros para cada objetivo estão representados na Tabela 7.

Tabela 6 – Resultados da etapa de validação para as capitais brasileiras.

Objetivo	R^2	MAE	RMSE
Infectados	0,9974	0,0370	0,0517
Óbitos	0,9976	0,0324	0,0483

Pela Tabela 7, o otimizador L-BFGS destacou-se como melhor dentre as 30 interações, para cada objetivo. Ambos os modelos apresentaram quantidades de neurônios na camada escondida próximas, 39 neurônios para o número de casos confirmados e 37 para o número de óbitos. Ademais, os outros hiperparâmetros diferem entre si, com destaque para as funções de ativação: logística para o primeiro objetivo, enquanto que a tangente hiperbólica apresentou melhor eficiência para o segundo caso. O tempo de treinamento para cada caso está representado na Tabela 8. As curvas estimadas por cada modelo

Tabela 7 – Hiperparâmetros para os melhores modelos referentes às capitais brasileiras. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.

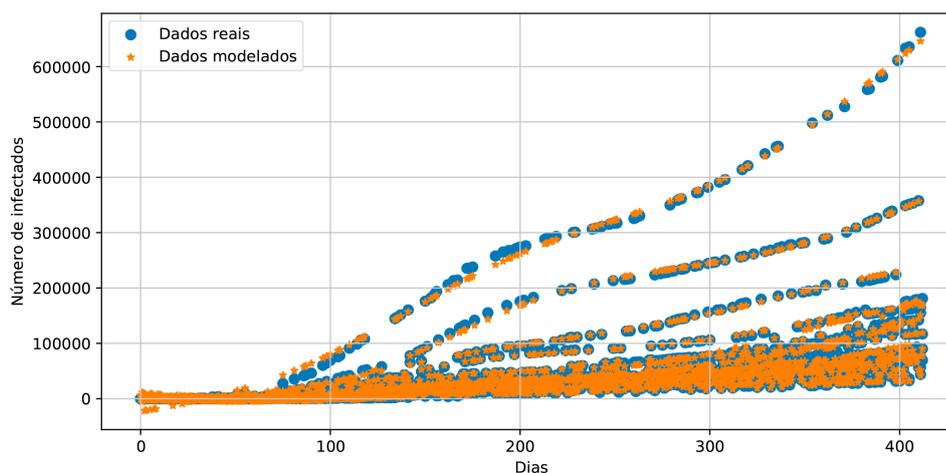
Hiperparâmetro	Modelo para n ^o de infectados	Modelo para n ^o de óbitos
Otimizador	LBFGS	LBFGS
Taxa de aprendizagem	0,01	0,5
Parâmetro de regularização	0,001	0,01
Tamanho das camadas escondidas	39	37
Função de ativação	logistic	tanh

estão ilustradas nas Figuras 6 e 7, nas quais é possível comparar os dados reais e aqueles previstos pelas arquiteturas de PMC.

Tabela 8 – Tempo de treinamento para cada objetivo, em minutos, nas capitais brasileiras, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.

Objetivo	Tempo de treinamento (min)
Número de infectados	4,163
Número de óbitos	5,203
Média	4,683

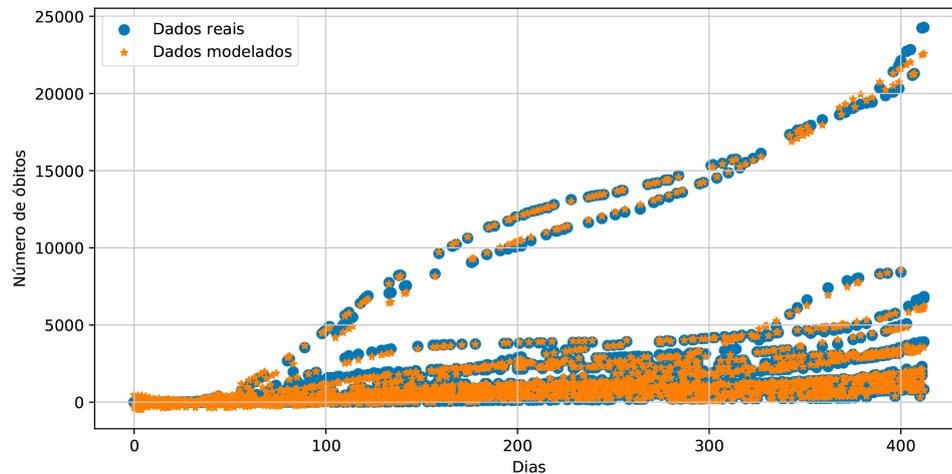
Figura 6 – Curva do número de pacientes infectados nas capitais brasileiras. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

Observa-se que, apesar do aumento do número de amostras (cerca de 27 vezes maior), o PMC foi capaz de modelar o comportamento da doença nos dois cenários com um aumento significativo do tempo de treinamento, quando comparado ao grupo de estudo da cidade de São Paulo, todavia, relativamente baixo.

Figura 7 – Curva do número de pacientes que vieram a óbito nas capitais brasileiras. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

4.3 Região Norte

Aqui estão representados os resultados referentes ao grupo de estudo da região Norte do Brasil. Essa região é constituída por 450 municípios, apresenta uma população de pouco mais de 15 milhões de habitantes, porém, consta com a menor densidade demográfica do país, de acordo com o censo de 2010 do IBGE (Instituto Brasileiro de Geografia e Estatística). Nesse cenário, *Perceptron* multicamadas foi alimentado com 142560 amostras durante a etapa de treinamento, a fim de encontrar a melhor combinação de hiperparâmetros. Os modelos selecionados foram validados com 35640 amostras novas.

O algoritmo de busca randomizada selecionou as redes com um coeficiente de determinação R^2 médio de 0,8123 e 0,9035 para o número total de pacientes infectados e para aqueles que vieram a óbito, respectivamente. As arquiteturas selecionadas foram validadas com os dados de teste. Os resultados dessa etapa estão presentes na Tabela 9. Os valores obtidos dos hiperparâmetros para cada objetivo estão representados na Tabela 11. O tempo de treinamento de todas as redes está representado na Tabela 10.

De princípio, observa-se o aumento significativo do tempo de treinamento como consequência do maior número de amostras. O otimizador L-BFGS continua prevalecendo sobre o *Adam*. Também, para ambos os modelos, tanto o parâmetro de regularização quanto a função de ativação foram iguais, 0.0001 e tangente hiperbólica (*tanh*).

Tabela 9 – Resultados da etapa de validação para a região Norte do Brasil.

Objetivo	R^2	MAE	RMSE
Infectados	0,7820	0,2034	0,4801
Óbitos	0,9024	0,1392	0,3166

Tabela 10 – Tempo de treinamento para cada objetivo, em minutos, na região Norte, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.

Objetivo	Tempo de treinamento (min)
Número de infectados	125,328
Número de óbitos	137,15
Média	131,239

Tabela 11 – Hiperparâmetros para os melhores modelos referentes à região Norte do Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.

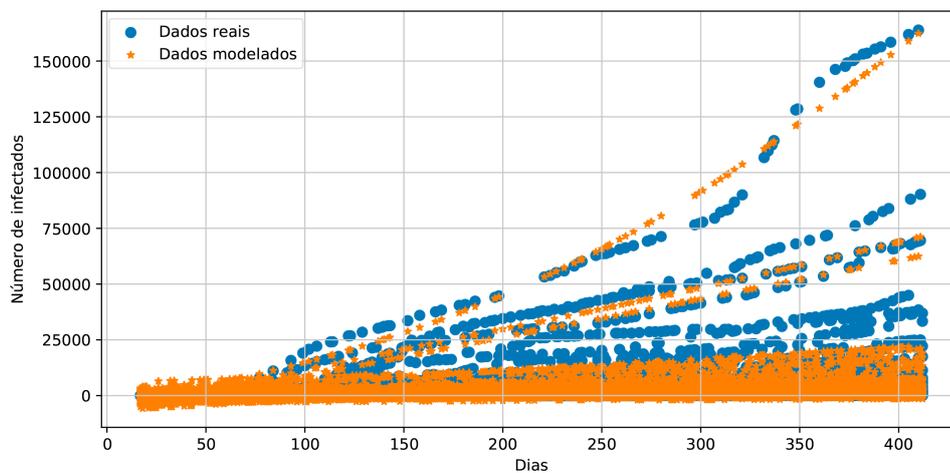
Hiperparâmetro	Modelo para n ^o de infectados	Modelo para n ^o de óbitos
Otimizador	LBFGS	LBFGS
Taxa de aprendizagem	0,001	0,01
Parâmetro de regularização	0,0001	0,0001
Tamanho das camadas escondidas	33	38
Função de ativação	tanh	tanh

Todavia, para este grupo de estudo, as redes selecionadas pelo algoritmo de busca randomizada apresentaram queda de desempenho. Apesar de mostrar um coeficiente de determinação R^2 relativamente alto, 0,7820 para o número de casos confirmados e 0,9024 para o número de óbitos, os parâmetros de erro, MAE e RMSE, observaram aumentos significativos, chegando a 50% para a raiz quadrada do erro-médio no primeiro objetivo, e cerca de 30% para o segundo. Estes resultados ficam evidenciados ao observar as curvas de comparação entre os dados reais e os modelados pela rede neural nas Figuras 8 e 9, nas quais há amostras em que os modelos não conseguiram aprender o comportamento, de modo que não houve sobreposição das curvas além da presença de valores negativos para o número de casos confirmados e de óbitos.

4.4 Região Centro Oeste

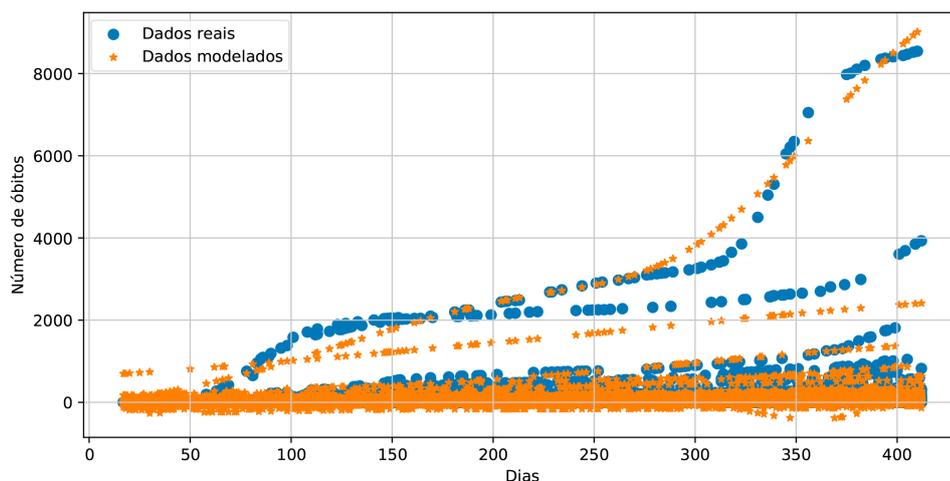
A região Centro Oeste, por sua vez, é composta por 466 municípios e possui uma população de pouco mais de 14 milhões de habitantes, correspondendo à segunda região menos populosa do Brasil, segundo dados do censo 2010 do IBGE. Para esse grupo de estudos, todas as redes PMC foram alimentadas com 150187 amostras durante a etapa de treinamento, as quais correspondem a dados de 412 dias desde o início da pandemia. Para validação dos resultados, os modelos selecionados foram testados com 37547 amostras novas, desconhecidas pelas redes treinadas.

Figura 8 – Curvas do número de pacientes infectados na região Norte do Brasil. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

Figura 9 – Curva do número de pacientes que vieram a óbito na região Norte do Brasil. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

Neste grupo de estudo, o coeficiente de determinação R^2 médio para as melhores redes, durante a etapa de treinamento, foram de 0,8642 para o modelo de número de infectados e 0,8403 para os óbitos. Ambas as arquiteturas foram validadas nos respectivos dados de teste. Os resultados dessa etapa estão presentes na Tabela 12. Os hiperparâmetros selecionados de cada modelo para cada objetivo estão representados na Tabela 13. O tempo de treinamento de todas as redes está representado na Tabela 14.

Tabela 12 – Resultados da etapa de validação para a região Centro Oeste do Brasil.

Objetivo	R^2	MAE	RMSE
Infectados	0,9152	0,1203	0,2827
Óbitos	0,9562	0,0889	0,2184

Tabela 13 – Hiperparâmetros para os melhores modelos referentes à região Centro Oeste do Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.

Hiperparâmetro	Modelo para n ^o de infectados	Modelo para n ^o de óbitos
Otimizador	LBFGS	LBFGS
Taxa de aprendizagem	0,0001	0,0001
Parâmetro de regularização	0,1	0,001
Tamanho das camadas escondidas	38	30
Função de ativação	tanh	tanh

Tabela 14 – Tempo de treinamento para cada objetivo, em minutos, na região Centro Oeste, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.

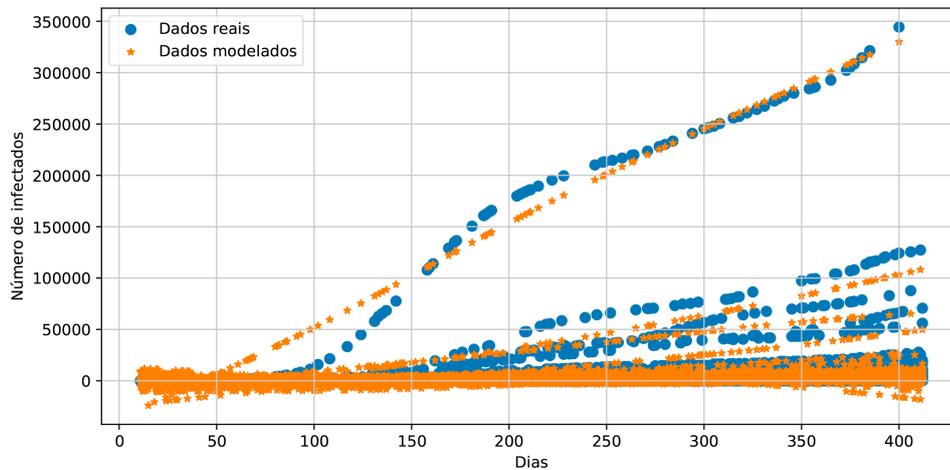
Objetivo	Tempo de treinamento (min)
Número de infectados	133,26
Número de óbitos	110,9
Média	122,08

Inicialmente, observa-se que o tempo se manteve perto dos valores referentes ao treinamento das redes para a região Norte, visto que a quantidade de amostras são próximas, de modo que tende a aumentar à medida que o número de amostras cresce.

Ademais, apesar do coeficiente de determinação médio em torno de 0,84 a 0,86 durante a etapa de treinamento, para ambos os objetivos, percebe-se um aumento na fase de validação: 0,9152 e 0,9562 para o número de casos confirmados e número de óbitos, respectivamente. Todavia, os valores para o erro médio absoluto e a raiz quadrada do erro-médio possuem valores consideráveis, conforme consta na Tabela 12. A ilustração desses resultados fica ainda mais evidente ao observar, nas Figuras 10 e 11, as curvas de comparação entre os dados reais e os modelados pelas redes neurais em cada caso analisado.

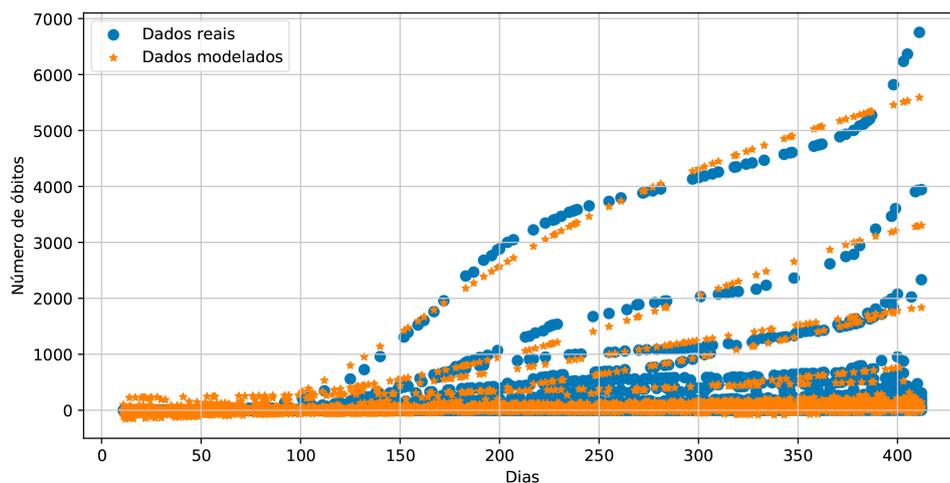
Portanto, apesar dos modelos possuírem um alto índice de representação dos dados reais, expresso pelo coeficiente de determinação R^2 , observa-se que há casos em que as diferenças entre os dados reais e os modelados é mais expressiva. Essa característica é melhor ilustrada para os casos confirmados da doença, enquanto que para aquele referente ao número de óbitos essa diferença é menor, visto que possui um R^2 próximo de 1, conforme também representado pelos valores de MAE e RMSE. Vale ressaltar, também, erros de ambos os modelos ao estimar valores negativos para cada objetivo.

Figura 10 – Curvas do número de pacientes infectados na região Centro Oeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

Figura 11 – Curva do número de pacientes que vieram a óbito na região Centro Oeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

4.5 Região Sul

Nesta seção serão abordados os resultados referentes ao grupo de estudo da região Sul do Brasil. Essa região compreende 1191 municípios e uma população de pouco mais de 27 milhões de habitantes, segundo dados do censo 2010 do IBGE. Todas as redes PMC foram alimentadas com 380167 amostras de entrada, referentes a 412 dias de pandemia. Os modelos selecionados foram validados a partir dos dados de teste, os quais correspondem a 95042 novas amostras.

Para este grupo, o coeficiente de determinação R^2 médio das melhores redes, durante a etapa de treinamento, foi de 0,6731 para o número de pacientes infectados pela doença, e 0,7456 para o número de pacientes que vieram a óbito. Aqui, já vale atentar-se

para a possibilidade de ambos os modelos não conseguirem aproximar bem as curvas dos dados reais, devido aos valores de R^2 na etapa de treino. Entretanto, a fim de confirmar essa afirmação, é necessário olhar para os resultados da etapa de validação, presentes na Tabela 15. Dessa forma, os hiperparâmetros selecionados para cada modelo estão representados na Tabela 16. O tempo de treinamento de todas as 150 redes, para cada objetivo, está presente na Tabela 17.

Tabela 15 – Resultados da etapa de validação para a região Sul do Brasil.

Objetivo	R^2	MAE	RMSE
Infectados	0,6158	0,2136	0,6631
Óbitos	0,7712	0,1664	0,4844

Tabela 16 – Hiperparâmetros para os melhores modelos referentes à região Sul do Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.

Hiperparâmetro	Modelo para n^o de infectados	Modelo para n^o de óbitos
Otimizador	LBFSGS	LBFSGS
Taxa de aprendizagem	0,01	0,001
Parâmetro de regularização	0,01	0,01
Tamanho das camadas escondidas	20	27
Função de ativação	tanh	logistic

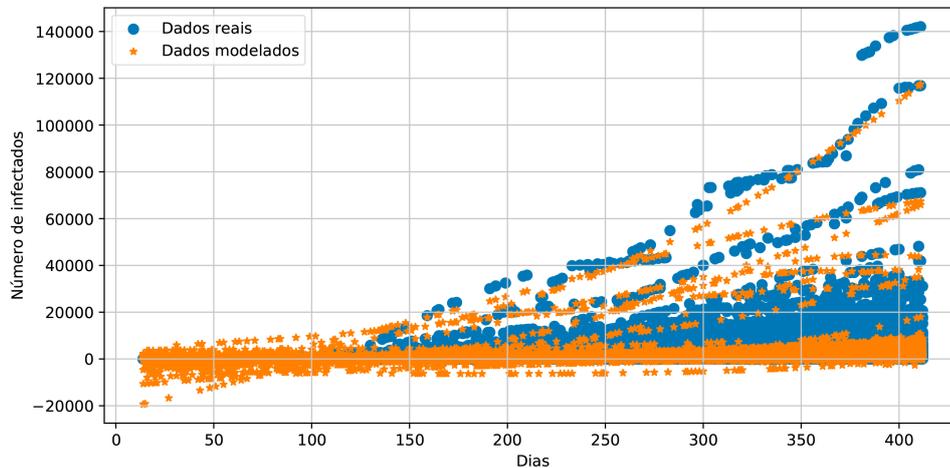
Em ambos os modelos, o otimizador L-BFGS apresentou aquele com maior desempenho frente ao *Adam*. Além disso, o parâmetro de regularização foi semelhante em ambos os casos. Todavia, percebe-se a diferença entre as taxas de aprendizado, o número de neurônios na camada escondida e a função de ativação. Vale chamar atenção para o número de neurônios na camada escondida, visto que nos modelos dos grupos de estudo anteriores esse resultado foi maior, de tal modo que os desempenhos das redes (R^2 , MAE e RMSE) também mostraram-se melhores. Ademais, não é possível notar diferença significativa entre as funções de ativação.

Com base nos resultados acima, percebe-se valores significativos para a raiz quadrada do erro-médio (RMSE), 0,6631 e 0,4844, para o número de infectados e número de óbitos, respectivamente. Isso implica que os valores previstos por ambos os modelos se encontram distantes das amostras reais. Somado a esse fato, valores altos para RMSE podem indicar a presença de *outliers*, uma vez que essas amostras possuem maior peso no cálculo do erro. Esses comportamentos estão ilustrados nas Figuras 12 e 13.

Com base nas curvas, observa-se que os modelos selecionados não modelaram satisfatoriamente as amostras reais, visto que há grandes espaços entre esses dados e aqueles previstos pelas redes, principalmente a partir do 200^o dia, para o número de infectados.

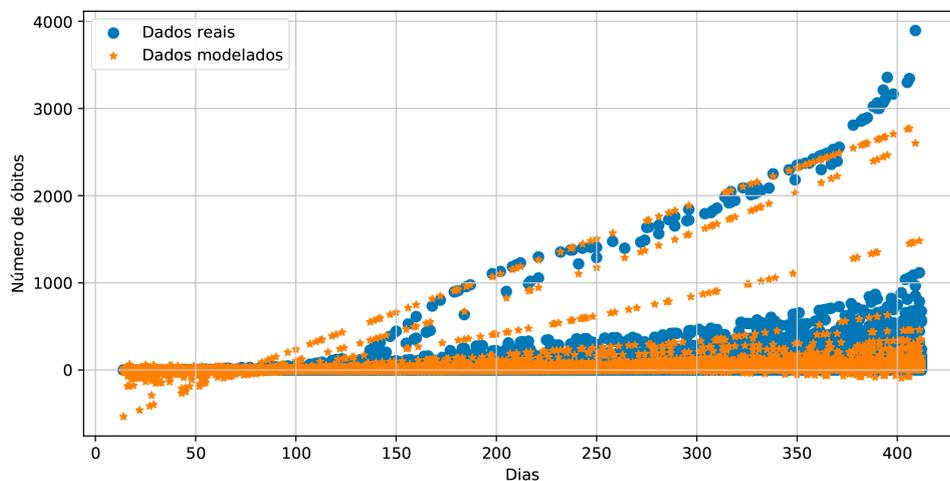
Ademais, nota-se a presença de valores negativos em ambos os cenários, ilustrando que os modelos não conseguiram extrair todas as características presentes nos dados.

Figura 12 – Curvas do número de pacientes infectados na região Sul do Brasil. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

Figura 13 – Curva do número de pacientes que vieram a óbito na região Sul do Brasil. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

Por fim, tem-se aumento do tempo de treinamento de todas as redes nos dois cenários, fato que é justificado pela quantidade maior de amostras (o dobro quando comparado à região Centro Oeste).

4.6 Região Sudeste

Esse grupo de estudo é marcado por ser o local no qual reportou-se o primeiro caso confirmado da COVID-19 no Brasil, no dia 25 de fevereiro de 2020. Também representa a

Tabela 17 – Tempo de treinamento para cada objetivo, em minutos, na região Sul, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.

Objetivo	Tempo de treinamento (min)
Número de infectados	384,5
Número de óbitos	228,33
Média	306,415

região geográfica brasileira com maior número acumulado de casos confirmados e óbitos da doença, segundo dados do Ministério da Saúde. Somado a isso, a região Sudeste conta com 1668 municípios, possui a maior densidade demográfica do país, além de ser a mais populosa, cerca de 80 milhões de habitantes, conforme o censo 2010 do IBGE.

Neste cenário, todas as redes PMC foram alimentadas com 467040 amostras de entradas, referentes a 412 dias de pandemia. Os modelos selecionados pelo algoritmo de busca randomizada foram validados a partir dos dados de teste, os quais correspondem a 116760 amostras, desconhecidas pelas redes. Durante a etapa de treinamento, o coeficiente de determinação R^2 médio, entre as partições da validação cruzada, foi de 0,8441 para o número de pacientes infectados, e 0,9083 para o número de óbitos. Os resultados da etapa de validação para os modelos selecionados estão representados na Tabela 18. A melhor combinação de hiperparâmetros para cada objetivo está presente na Tabela 19. Por fim, o tempo de treinamento de todas as 150 redes neurais, para cada cenário, está representado na Tabela 20.

Tabela 18 – Resultados da etapa de validação para a região Sudeste do Brasil.

Objetivo	R^2	MAE	RMSE
Infectados	0,8653	0,1113	0,3578
Óbitos	0,9163	0,0809	0,2852

Tabela 19 – Hiperparâmetros para os melhores modelos referentes à região Sudeste do Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.

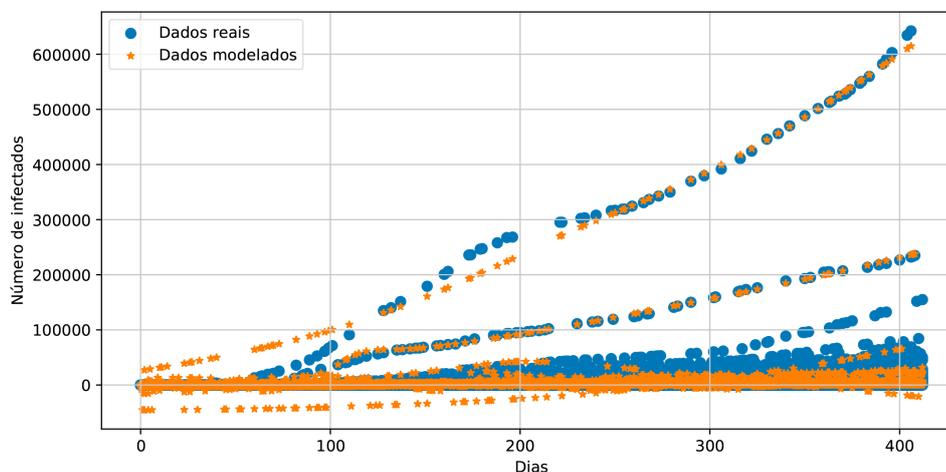
Hiperparâmetro	Modelo para n^o de infectados	Modelo para n^o de óbitos
Otimizador	LBFSGS	LBFSGS
Taxa de aprendizagem	0,01	0,001
Parâmetro de regularização	0,0001	0,001
Tamanho das camadas escondidas	32	29
Função de ativação	tanh	logistic

Novamente, o otimizador L-BFGS obteve resultados melhores frente ao *Adam*. Entretanto, todos os outros valores diferem para cada objetivo. Nota-se que, para o segundo cenário, o modelo encontrado possui em sua camada escondida 29 neurônios, todavia, seus resultados mostraram-se superiores comparados ao primeiro cenário, com 32 neurônios na camada escondida. Para esse grupo de estudo, o tempo de treinamento aumentou quando comparado ao grupo anterior, devido ao aumento do número de amostras. Os resultados das Tabelas 18 e 19 estão melhor ilustrados nas curvas de cada objetivo, presentes nas Figuras 14 e 15.

Tabela 20 – Tempo de treinamento para cada objetivo, em minutos, na região Sudeste, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.

Objetivo	Tempo de treinamento (min)
Número de infectados	335,23
Número de óbitos	330,66
Média	332,945

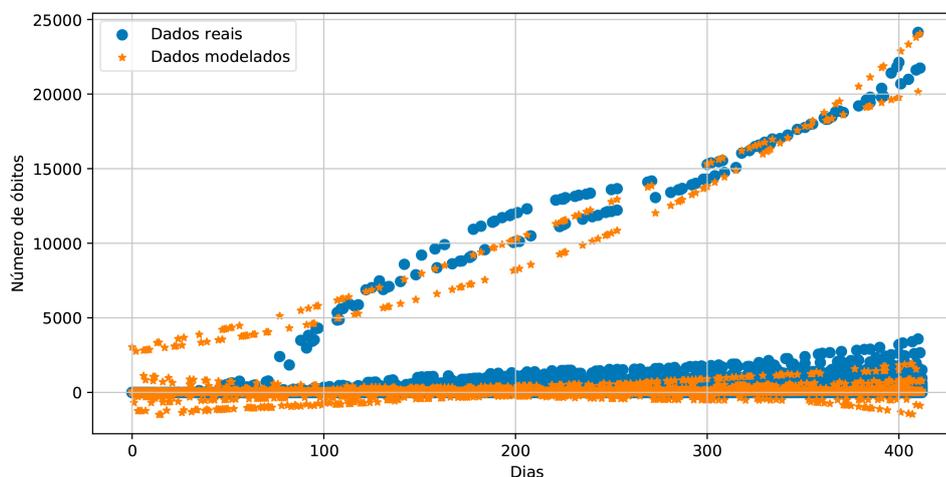
Figura 14 – Curvas do número de pacientes infectados na região Sudeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

A partir das curvas acima, nota-se que ambos os modelos representam os valores reais com erros significativos, de modo que representar o comportamento preciso da doença em cada localidade não foi plenamente satisfeito. Nota-se, também, a presença de regiões sem sobreposição das curvas. Somado a isso, observa-se a presença de valores negativos que foram retratados por ambos os modelos. Dessa forma, apesar dos valores consideráveis para R^2 , presentes na Tabela 18, ainda há erros significativos por parte dos modelos.

Figura 15 – Curva do número de pacientes que vieram a óbito na região Sudeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

4.7 Região Nordeste

Este grupo de estudo compreende um total de 578385 amostras, as quais serviram de entrada para as redes neurais, durante a etapa de treinamento. As redes oriundas das melhores combinações de hiperparâmetros para cada objetivo, foram validadas com os dados de teste, os quais correspondem a cerca de 144597 amostras.

Os modelos selecionados pelo algoritmo de busca randomizada obtiveram, durante a etapa de treinamento, um coeficiente de determinação R^2 médio, entre as partições da validação cruzada, de 0,477 para o número de pacientes infectados, e 0,6096 para o número de óbitos. Os resultados da etapa de validação para os modelos selecionados estão presentes na Tabela 21, enquanto que a combinação de hiperparâmetros para cada objetivo está representada na Tabela 22. Por fim, o tempo de treinamento de todas as 150 arquiteturas de PMC, para cada cenário, está representado na Tabela 23.

Tabela 21 – Resultados da etapa de validação para a região Nordeste do Brasil.

Objetivo	R^2	MAE	RMSE
Infectados	0,4673	0,2406	0,6944
Óbitos	0,4901	0,2364	0,6859

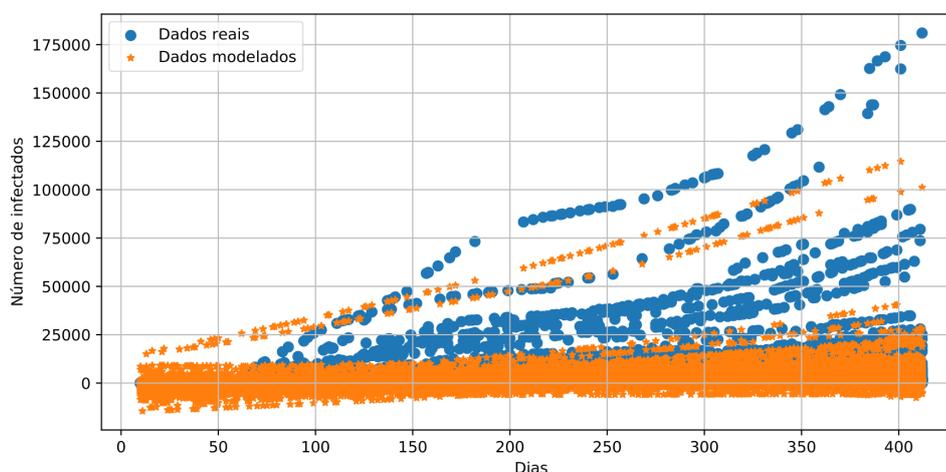
Aqui, já é possível observar a partir dos dados das tabelas acima que ambos os modelos possuem valores semelhantes de hiperparâmetros, divergindo quanto ao parâmetro de regularização e ao número de neurônios na camada escondida. Todavia, os resultados da etapa de validação mostram que ambas arquiteturas não foram suficientes para aprender com eficiência o comportamento da doença nos dois objetivos propostos, visto que possuem um R^2 inferior a 50%, ou seja, os modelos propostos só foram capazes de representar cerca

Tabela 22 – Hiperparâmetros para os melhores modelos referentes à região Nordeste do Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.

Hiperparâmetro	Modelo para n ^o de infectados	Modelo para n ^o de óbitos
Otimizador	LBFSGS	LBFSGS
Taxa de aprendizagem	0,0001	0,0001
Parâmetro de regularização	0,0001	0,1
Tamanho das camadas escondidas	36	30
Função de ativação	tanh	tanh

de metade dos dados reais. Os valores obtidos para MAE e RMSE reforçam ainda mais essa falha. O comportamento de ambos os modelos podem ser observados nas Figuras 16 e 17 a seguir. Vale ressaltar, também, o aumento do tempo de treinamento, principalmente em relação ao modelo do número de pacientes infectados, uma vez que aumentou-se o número de amostras nesse cenário.

Figura 16 – Curvas do número de pacientes infectados na região Nordeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.



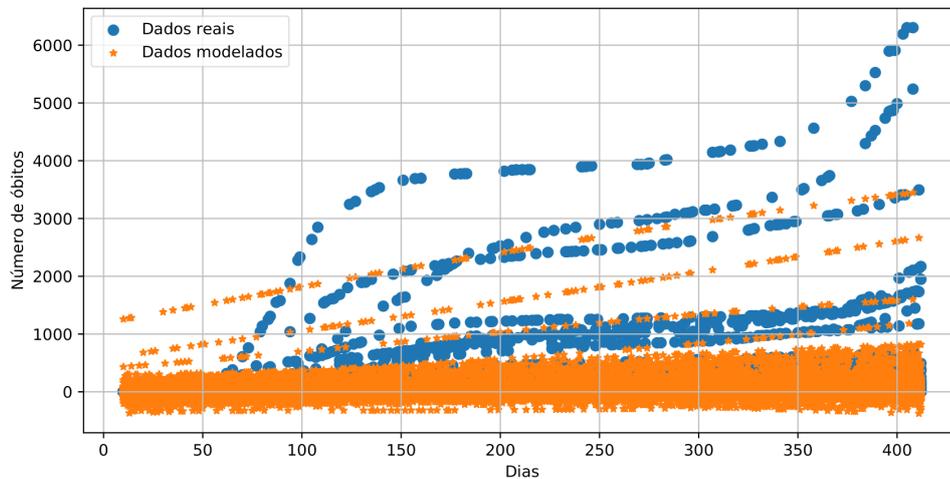
Fonte: Elaborado pelo autor.

A partir da observação das ilustrações acima, é possível notar que ambos os modelos sofrem ao tentar modelar o comportamento da doença nos dois cenários. Isso é retratado pela presença de valores previstos pelos modelos abaixo do eixo X, ou seja, números negativos de pacientes.

4.8 Brasil

Por fim, tem-se o grupo de estudo do Brasil, o qual compreende todos os 5568 municípios somado do Distrito Federal, com uma população de pouco mais de 190 milhões

Figura 17 – Curva do número de pacientes que vieram a óbito na região Nordeste do Brasil. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

Tabela 23 – Tempo de treinamento para cada objetivo, em minutos, na região Nordeste, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.

Objetivo	Tempo de treinamento (min)
Número de infectados	551,33
Número de óbitos	307,46
Média	429,39

de habitantes, segundo censo 2010 do IBGE. Para esse cenário, todas as redes PMC foram alimentadas com 1840328 amostras de entrada, referentes a 412 dias de pandemia. Os modelos selecionados foram validados a partir dos dados de teste, os quais correspondem a 460082 novas amostras.

Para este grupo, o coeficiente de determinação R^2 médio, durante a etapa de treinamento, foi de 0,5949 para o número de pacientes infectados, e 0,3319 para o número de pacientes que vieram a óbito no país em decorrente da COVID-19. Os resultados referentes à etapa de validação estão presentes na Tabela 24, enquanto que as combinações de hiperparâmetros para os dois objetivos estão representadas na Tabela 25. Por fim, o tempo de treinamento para cada cenário pode ser observado na Tabela 26.

Tabela 24 – Resultados da etapa de validação para o Brasil.

Objetivo	R^2	MAE	RMSE
Infectados	0,5832	0,1591	0,6387
Óbitos	0,3236	0,1671	0,8614

Tabela 25 – Hiperparâmetros para os melhores modelos referentes ao Brasil. Cada coluna representa um modelo para cada objetivo - número de infectados, número de óbitos.

Hiperparâmetro	Modelo para n ^o de infectados	Modelo para n ^o de óbitos
Otimizador	LBFGS	LBFGS
Taxa de aprendizagem	0,01	0,01
Parâmetro de regularização	0,01	0,5
Tamanho das camadas escondidas	18	18
Função de ativação	tanh	tanh

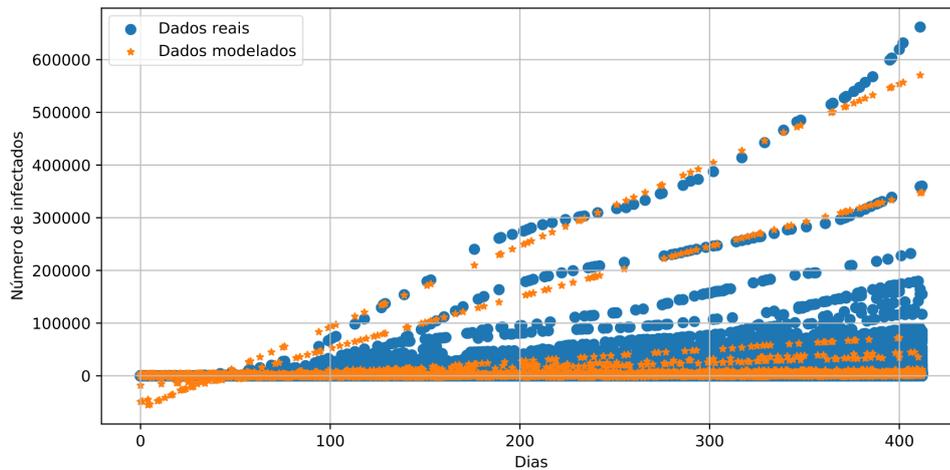
A partir dos resultados acima, percebe-se que as arquiteturas de PMC escolhidas possuem valores semelhantes de hiperparâmetros, diferindo apenas quanto ao parâmetro de regularização. Todavia, tiveram um R^2 médio inferior a 0,60. O RMSE para ambos os casos é superior a 0,60, chegando a 0,86 para o número de óbitos. Para este último cenário, o coeficiente de determinação R^2 foi de 0,32 na etapa de validação, isso mostra que o modelo não foi capaz de "explicar" os dados reais, ademais, o valor de RMSE de 0,86 infere a presença de valores muito acima da média do conjunto de dados, de tal modo que o modelo teve dificuldades em interpretar essas características, conforme observado na Figura 19. Além disso, a partir da Figura 18 observa-se que apesar do R^2 ter sido maior para esse cenário, o PMC sofre ao prever valores negativos para o número de pacientes infectados. Somado a esse fato, também é nítido a presença de regiões nas quais não houve sobreposições das curvas de dados reais e dados modelados, conforme previsto pelos valores de erros expressos na Tabela 24.

Tabela 26 – Tempo de treinamento para cada objetivo, em minutos, no Brasil, utilizando o algoritmo de validação cruzada com 5 partições e busca randomizada com 30 interações, totalizando 150 redes neurais artificiais treinadas.

Objetivo	Tempo de treinamento (min)
Número de infectados	863,35
Número de óbitos	838,05
Média	850,7

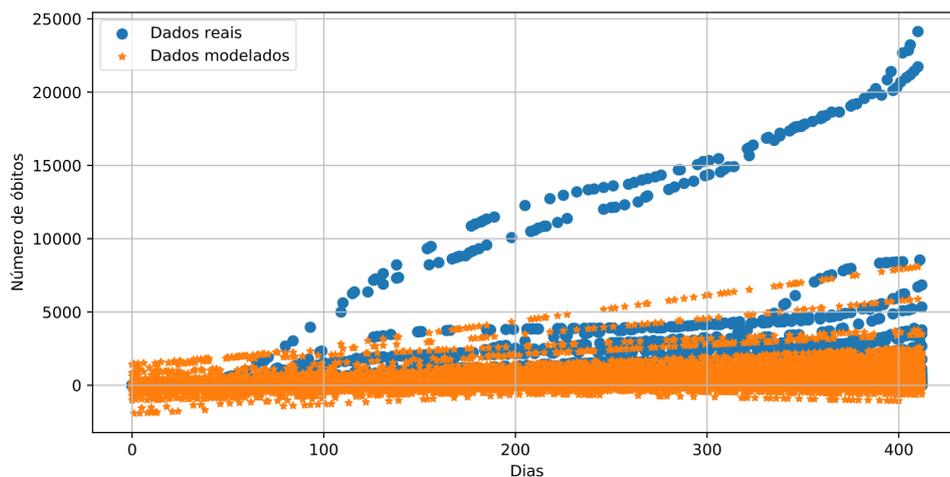
Por fim, tem-se valores elevados para o tempo de treinamento em cada objetivo, os quais são justificáveis pela enorme quantidade de amostras, cerca de 1,8 milhão para essa etapa. Todavia, esse tempo de processamento produziu modelos que obtiveram alto índice de erro.

Figura 18 – Curvas do número de pacientes infectados no Brasil. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

Figura 19 – Curva do número de pacientes que vieram a óbito no Brasil. Comparação entre os dados reais e os modelados pelo PMC.



Fonte: Elaborado pelo autor.

4.9 Análise Geral

A partir dos resultados anteriores, pode-se perceber diferenças entre os modelos treinados para cada grupo de estudo e objetivo. O coeficiente de determinação R^2 durante a etapa de validação variou de 0,32, como no caso do Brasil, a 0,99 para os cenários de São Paulo/SP e capitais brasileiras, conforme mostrado na Tabela 27. Dentre as quatro funções de ativação disponíveis durante o processo de treinamento e otimização de hiperparâmetros, duas tiveram maior destaque, a tangente hiperbólica (*tanh*) e a função logística (*logistic*), sendo as mais escolhidas para os melhores modelos. Todavia, a função linear retificada (ReLU) apareceu uma única vez dentre as melhores combinações. Um ponto que vale

ressaltar refere-se à quantidade de neurônios na camada oculta. Pelas tabelas (4, 7, 11, 13, 16, 19, 22 e 25), apresentadas nas seções anteriores, os modelos que, em sua maioria, apresentaram maior eficiência em representar os dados reais são aqueles que possuem dentre 30 a 40 neurônios nessa camada. Ademais, é importante frisar que em todos os cenários, o otimizador de pesos sinápticos escolhido pelos algoritmos foi o L-BFGS. Por fim, observa-se o incremento do tempo de treinamento em cada cenário proporcional à grandeza do banco de dados utilizado, chegando a cerca de 13 horas de treinamento para o modelo do número de casos confirmados da doença no grupo de estudo do Brasil.

Tabela 27 – Coeficiente de determinação R^2 de cada objetivo em cada grupo de estudo.

Grupo de estudo	Nº de infectados	Nº de óbitos
São Paulo/SP	0,9999	0,9989
Capitais brasileiras	0,9974	0,9976
Região Norte	0,7820	0,9024
Região Centro Oeste	0,9152	0,9562
Região Sul	0,6158	0,7712
Região Sudeste	0,9653	0,9163
Região Nordeste	0,4673	0,4901
Brasil	0,5832	0,3236

5 Considerações Finais

Este trabalho propôs a utilização de redes PMC com uma única camada oculta para modelar, separadamente, o número máximo de pacientes infectados e óbitos por COVID-19 no Brasil. A metodologia escolhida foi aplicada a 8 grupos de estudo, os quais compreendem São Paulo/SP, as capitais brasileiras somadas do Distrito Federal, as cinco regiões geográficas do Brasil e, por fim, os dados englobando todo o país. Dividir o trabalho nesses grupos permite avaliar a capacidade de generalização do PMC como um aproximador de funções. A fim de obter a rede neural com maior eficiência aplicou-se o algoritmo de busca randomizada de modo a encontrar a combinação de hiperparâmetros da rede que tivesse o melhor desempenho durante a aplicação do algoritmo de validação cruzada. Ao todo foram treinadas 150 redes para cada objetivo (número de infectados e número de óbitos) e em cada grupo. Dentre os resultados, observou-se um coeficiente de determinação R^2 em torno de 0,99 para os dois primeiros grupos de estudo. Para a região Centro Oeste e Sudeste, esse valor fica compreendido entre 0,91 e 0,96 para os dois cenários. (CAR et al., 2020) obteve valores semelhantes em seus resultados, porém para arquiteturas mais complexas, envolvendo quatro camadas ocultas e um conjunto de dados referente a 406 localidades e 51 dias de pandemia, totalizando 20706 amostras. Por outro lado, à medida que a metodologia foi aplicada a regiões de maior abrangência, aumentando assim a complexidade dos dados, notou-se uma queda na eficiência dos modelos, chegando a um R^2 em torno de 0,32 para o número de óbitos do Brasil. Com isso, conclui-se que é possível modelar a propagação da COVID-19 tomando-se a latitude, longitude e o número de dias desde o primeiro caso reportado da doença como entradas de um PMC, de modo que a metodologia proposta pode desempenhar excelentes resultados quando trata-se de regiões menores como, cidades, capitais e até mesmo estados. Porém, à medida que a complexidade dos dados aumenta, observa-se queda do desempenho das redes, de modo que a metodologia de PMC com apenas uma única camada oculta pode não ser suficiente para representar os dados reais. A partir disso, o próximo passo é desenvolver modelos de redes neurais mais complexos através da aplicação de ferramentas de *Deep Learning*, uma vez que possibilitaria propor e validar diferentes metodologias para o problema em um menor intervalo de tempo.

Referências

- BENGIO, Y. Practical recommendations for gradient-based training of deep architectures. *Arxiv*, 06 2012. Citado 2 vezes nas páginas 19 e 20.
- BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, v. 13, p. 281–305, 2012. Disponível em: <<http://dblp.uni-trier.de/db/journals/jmlr/jmlr13.html#BergstraB12>>. Citado na página 26.
- BISHOP, C. *Neural networks for pattern recognition*. [S.l.]: Oxford University Press, USA, 1995. Citado na página 18.
- BRITTON, A. et al. Effectiveness of the pfizer-BioNTech COVID-19 vaccine among residents of two skilled nursing facilities experiencing COVID-19 outbreaks — connecticut, december 2020–february 2021. *MMWR. Morbidity and Mortality Weekly Report*, Centers for Disease Control MMWR Office, v. 70, n. 11, p. 396–401, mar. 2021. Disponível em: <<https://doi.org/10.15585/mmwr.mm7011e3>>. Citado na página 14.
- BUITINCK, L. et al. API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. [S.l.: s.n.], 2013. p. 108–122. Citado 2 vezes nas páginas 21 e 25.
- CANDIDO, D. S. et al. Evolution and epidemic spread of sars-cov-2 in brazil. *Science*, American Association for the Advancement of Science, v. 369, n. 6508, p. 1255–1260, 2020. ISSN 0036-8075. Disponível em: <<https://science.sciencemag.org/content/369/6508/1255>>. Citado na página 14.
- CAR, Z. et al. Modeling the spread of covid-19 infection using a multilayer perceptron. *Computational and Mathematical Methods in Medicine*, v. 2020, p. 1–10, 05 2020. Citado 6 vezes nas páginas 14, 15, 20, 24, 27 e 47.
- COTA, W. Monitoring the number of COVID-19 cases and deaths in brazil at municipal and federative units level. *SciELOPreprints:362*, FapUNIFESP (SciELO), maio 2020. Disponível em: <<https://doi.org/10.1590/scielopreprints.362>>. Citado 2 vezes nas páginas 14 e 22.
- Diaz, G. I. et al. An effective algorithm for hyperparameter optimization of neural networks. *IBM Journal of Research and Development*, v. 61, n. 4/5, p. 9:1–9:11, 2017. Citado na página 19.
- DZIERŻAK, R. Comparison of the influence of standardization and normalization of data on the effectiveness of spongy tissue texture classification. *Informatyka, Automatyka, Pomiar w Gospodarce i Ochronie Środowiska*, v. 9, p. 66–69, 09 2019. Citado na página 17.
- FEUER, W. *South America is a ‘new epicenter’ of the coronavirus pandemic, WHO says*. 2020. Disponível em: <<https://www.cnbc.com/2020/05/22/south-america-is-a-new-epicenter-of-the-coronavirus-pandemic-who-says.html>>. Citado na página 13.

- GARDNER, M.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, v. 32, n. 14, p. 2627–2636, 1998. ISSN 1352-2310. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1352231097004470>>. Citado 3 vezes nas páginas 17, 18 e 19.
- GUAN, W.-J. et al. Clinical characteristics of coronavirus disease 2019 in china. *New England Journal of Medicine*, v. 382, 02 2020. Citado na página 13.
- HIPPERT, H.; PEDREIRA, C.; SOUZA, R. Neural networks for short-term load forecasting: A review and evaluation. *Power Systems, IEEE Transactions on*, v. 16, p. 44 – 55, 03 2001. Citado na página 18.
- IBRAHIM, S. Performance evaluation of multi-layer perceptron (mlp) and radial basis function (rbf): Covid-19 spread and death contributing factors. *International Journal of Advanced Trends in Computer Science and Engineering*, v. 9, p. 625–631, 09 2020. Citado na página 15.
- JAIN, S.; SHUKLA, S.; WADHVANI, R. Dynamic selection of normalization techniques using data complexity measures. *Expert Systems with Applications*, v. 106, p. 252–262, 2018. ISSN 0957-4174. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S095741741830232X>>. Citado na página 17.
- JAYALAKSHMI, T.; A., S. Statistical normalization and back propagation for classification. *International Journal Computer Theory Engineering (IJCTE)*, v. 3, p. 89–93, 01 2011. Citado na página 24.
- KARAKI, Y.; IVANOV, N. Hyperparameters of multilayer perceptron with normal distributed weights. *Pattern Recognition and Image Analysis*, v. 30, p. 170–173, 04 2020. Citado na página 19.
- KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. In: BENGIO, Y.; LECUN, Y. (Ed.). *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. [s.n.], 2015. Disponível em: <<http://arxiv.org/abs/1412.6980>>. Citado na página 21.
- KNOLL, M. D.; WONODI, C. Oxford–AstraZeneca COVID-19 vaccine efficacy. *The Lancet*, Elsevier BV, v. 397, n. 10269, p. 72–74, jan. 2021. Disponível em: <[https://doi.org/10.1016/s0140-6736\(20\)32623-4](https://doi.org/10.1016/s0140-6736(20)32623-4)>. Citado na página 14.
- LAROCHELLE, H. et al. Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, v. 1, p. 140, 01 2009. Citado na página 20.
- LI, Q. et al. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, v. 382, n. 13, p. 1199–1207, 2020. PMID: 31995857. Disponível em: <<https://doi.org/10.1056/NEJMoa2001316>>. Citado na página 13.
- MAIER, H.; DANDY, G. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environmental Modelling and Software*, v. 15, p. 101–124, 01 2000. Citado na página 20.

- MASSUDA, A. et al. The brazilian health system at crossroads: progress, crisis and resilience. *BMJ global health*, BMJ Publishing Group, v. 3, n. 4, p. e000829–e000829, Jul 2018. ISSN 2059-7908. 29997906[pmid]. Disponível em: <<https://pubmed.ncbi.nlm.nih.gov/29997906>>. Citado na página 14.
- MENARD, S. Coefficients of determination for multiple logistic regression analysis. *American Statistician - AMER STATIST*, v. 54, p. 17–24, 02 2000. Citado na página 27.
- MOHAMAD, I.; USMAN, D. Standardization and its effects on k-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, v. 6, p. 3299–3303, 2013. Citado na página 17.
- MOLLALO, A.; RIVERA, K. M.; VAHEDI, B. Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental united states. *International Journal of Environmental Research and Public Health*, MDPI AG, v. 17, n. 12, p. 4204, jun. 2020. Disponível em: <<https://doi.org/10.3390/ijerph17124204>>. Citado na página 15.
- NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika*, v. 78, n. 3, p. 691–692, 09 1991. ISSN 0006-3444. Disponível em: <<https://doi.org/10.1093/biomet/78.3.691>>. Citado na página 26.
- NAJARAN, M. Applications of artificial intelligence in battling against covid-19: A literature review. *Chaos, Solitons Fractals*, v. 142, p. 110338, 10 2020. Citado na página 15.
- PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 3 vezes nas páginas 21, 24 e 25.
- Popa, C. Quasi-newton learning methods for complex-valued neural networks. In: *2015 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2015. p. 1–8. Citado na página 21.
- RANZANI, O. T. et al. Characterisation of the first 250,000 hospital admissions for covid-19 in brazil: a retrospective analysis of nationwide data. *The Lancet Respiratory Medicine*, Elsevier, v. 9, n. 4, p. 407–418, Apr 2021. ISSN 2213-2600. Disponível em: <[https://doi.org/10.1016/S2213-2600\(20\)30560-9](https://doi.org/10.1016/S2213-2600(20)30560-9)>. Citado 2 vezes nas páginas 13 e 14.
- RIZK-ALLAH, R.; HASSANIEN, A. E. Covid-19 forecasting based on an improved interior search algorithm and multi-layer feed forward neural network. 04 2020. Citado na página 15.
- ROTHAN, H. A.; BYRAREDDY, S. N. The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak. *Journal of Autoimmunity*, v. 109, p. 102433, 2020. ISSN 0896-8411. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0896841120300469>>. Citado na página 13.
- RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning Representations by Back-propagating Errors. *Nature*, v. 323, n. 6088, p. 533–536, 1986. Disponível em: <<http://www.nature.com/articles/323533a0>>. Citado na página 18.
- SAVI, P. V.; SAVI, M. A.; BORGES, B. A mathematical description of the dynamics of coronavirus disease 2019 (covid-19): A case study of brazil. *Computational and Mathematical Methods in Medicine*, Hindawi, v. 2020, p. 9017157, Sep 2020. ISSN

1748-670X. Disponível em: <<https://doi.org/10.1155/2020/9017157>>. Citado na página 14.

Schilling, N. et al. Joint model choice and hyperparameter optimization with factorized multilayer perceptrons. In: *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*. [S.l.: s.n.], 2015. p. 72–79. Citado na página 19.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. *Redes Neurais Artificiais para Engenharia e Ciências Aplicadas: Fundamentos Teóricos e Aspectos Práticos*. 2ª edição. ed. São Paulo: Artliber, 2016. Citado 6 vezes nas páginas 18, 19, 20, 21, 24 e 25.

Sola, J.; Sevilla, J. Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science*, v. 44, n. 3, p. 1464–1468, 1997. Citado na página 24.

VABALAS, A. et al. Machine learning algorithm validation with a limited sample size. *PLOS ONE*, Public Library of Science, v. 14, n. 11, p. 1–20, 11 2019. Disponível em: <<https://doi.org/10.1371/journal.pone.0224365>>. Citado na página 25.

WANG, W.; TANG, J.; WEI, F. Updated understanding of the outbreak of 2019 novel coronavirus (2019-ncov) in wuhan, china. *Journal of Medical Virology*, v. 92, n. 4, p. 441–447, 2020. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.25689>>. Citado na página 13.

WORLDOMETERS. *COVID-19 CORONAVIRUS PANDEMIC*. 2021. Disponível em: <<https://www.worldometers.info/coronavirus/>>. Citado na página 13.

WU, R. et al. A l-bfgs based learning algorithm for complex-valued feedforward neural networks. *Neural Process. Lett.*, Kluwer Academic Publishers, USA, v. 47, n. 3, p. 1271–1284, jun. 2018. ISSN 1370-4621. Disponível em: <<https://doi.org/10.1007/s11063-017-9692-5>>. Citado na página 21.

WU, Z. et al. Safety, tolerability, and immunogenicity of an inactivated SARS-CoV-2 vaccine (CoronaVac) in healthy adults aged 60 years and older: a randomised, double-blind, placebo-controlled, phase 1/2 clinical trial. *The Lancet Infectious Diseases*, Elsevier BV, fev. 2021. Disponível em: <[https://doi.org/10.1016/s1473-3099\(20\)30987-7](https://doi.org/10.1016/s1473-3099(20)30987-7)>. Citado na página 14.

ZHAO, S. et al. Estimating the unreported number of novel coronavirus (2019-ncov) cases in china in the first half of january 2020: A data-driven modelling analysis of the early outbreak. *Journal of Clinical Medicine*, v. 9, p. 388, 02 2020. Citado na página 15.