

Data da publicação 23 Jun. 2020, data da versão atual 23 Abr. 2020.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

# Aprendizado de Máquina Aplicado à Análise de Dados do COVID-19

**PAULO GIOVANI<sup>1</sup>, FILIPE VERRI<sup>2</sup>**

<sup>1</sup>Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Campos do Jordão, Brasil (e-mail: paulogiovani@ifsp.edu.br)

<sup>2</sup>Divisão de Ciência da Computação (IEC), Instituto de Tecnologia Aeronáutica (ITA), São José dos Campos, Brasil (e-mail: verri@ita.br)

Endereço para correspondência: Paulo Giovani (e-mail: paulogiovani@ifsp.edu.br).

## RESUMO

O surgimento do coronavírus na China no final de 2019 foi responsável por uma grande mudança no comportamento global. Com o aumento do número de casos, as autoridades de saúde se viram obrigadas a decretar estado de pandemia. O estudo da evolução do número de casos é importante para que ações preventivas possam ser tomadas, visando minimizar os prejuízos para a sociedade. Dessa forma, o presente artigo tem por objetivo realizar um estudo a respeito da aplicação de técnicas de aprendizado de máquina para realizar a predição do número de casos confirmados de coronavírus ao redor do mundo. O resultado obtido permitiu verificar a importância de um bom conjunto de dados bem como qual intervalo de tempo é o mais indicado para que tal análise possa ser realizada de forma mais adequada.

**PALAVRAS-CHAVE** COVID-19, Análise de Dados, Aprendizado de Máquina, Séries Temporais, Redes Neurais, Teorema de Takens.

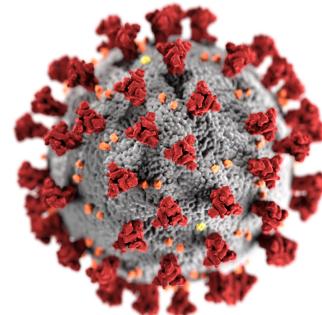
## I. INTRODUÇÃO

DESDE a descoberta dos vírus, no final do século passado, sempre houve muitas especulações a respeito de sua natureza. Tais questões envolvem, por exemplo, o fato deles serem considerados agentes extrínsecos, microrganismos vivos ou ainda, organismos não organizados intrínsecos à sua origem [1].

Os vírus são considerados organismos microscópicos, que precisam infectar um hospedeiro vivo para se reproduzir. Eles podem variar em sua complexidade, mas, geralmente, são compostos de material genético como DNA ou RNA, cercados por uma camada protetora de proteínas, a qual não podem replicar [2].

Existem inúmeros tipos de vírus e eles podem infectar tanto humanos como animais, sendo que alguns deles podem causar sérias complicações no organismo, inclusive, colo- cando vidas em risco. Eles podem ser classificados de acordo com as categorias descritas pelo Sistema de Baltimore<sup>1</sup>, a qual é baseada na síntese do RNA mensageiro, sendo que os seres humanos podem ser infectados por vírus existentes em todas essas categorias.

O Sistema de Baltimore classifica os diferentes tipos de vírus, agrupando-os em famílias, de acordo com o tipo de

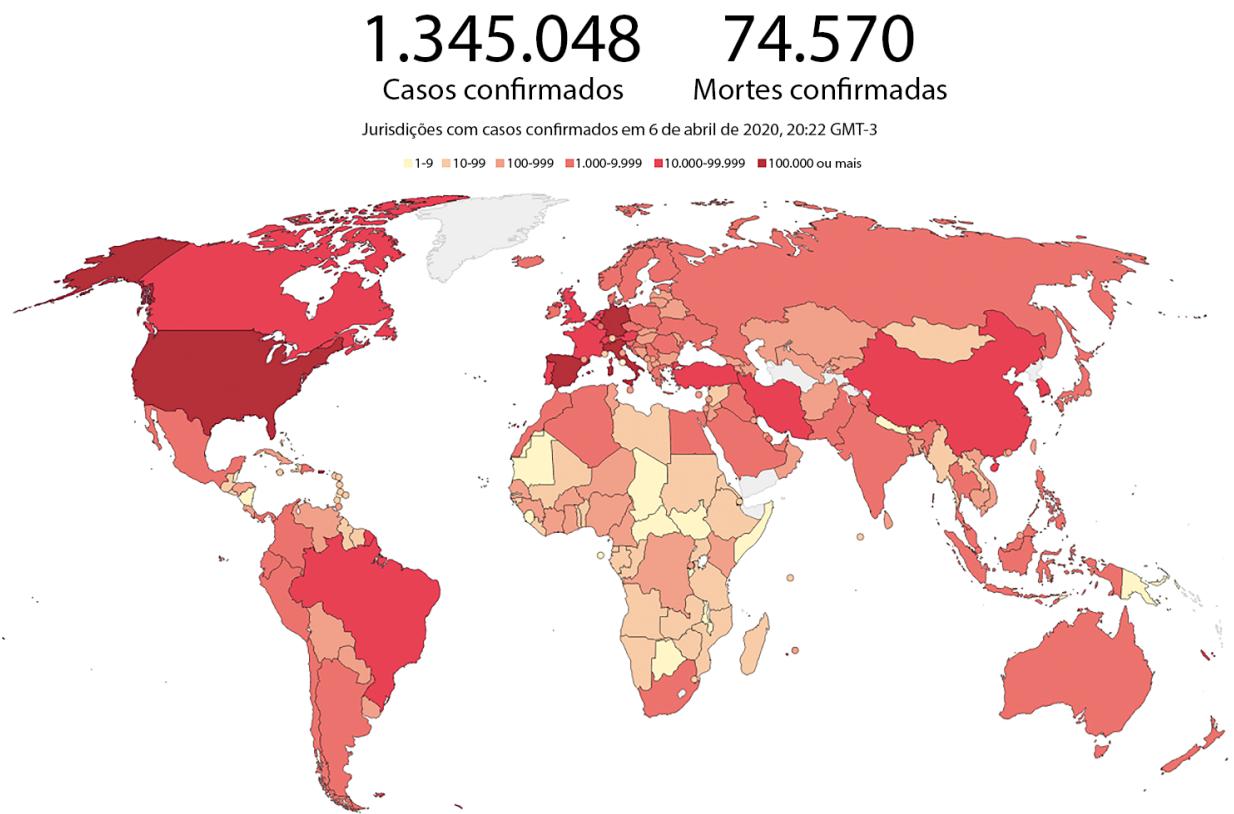


**Figura 1.** Ilustração da morfologia apresentada pelo coronavírus, criada pelo Centers for Disease Control and Prevention, onde é possível observar as espinhas vermelhas responsáveis por conferir a aparência de uma coroa, quando observado por meio de um microscópio eletrônico.

seu material genético (DNA, RNA, cadeia dupla e cadeia simples) e seu método de replicação<sup>2</sup>. Dentro da classificação definida pelo Sistema de Baltimore temos a categoria IV, a qual inclui os vírus do tipo (+)ssRNA. Esses tipos de vírus possuem material genético constituído por RNA de cadeia simples e senso positivo, e são considerados os mais abundantes do planeta. Dentre a sua família, temos a presença

<sup>1</sup>C. Shaffer. *O Sistema de Classificação de Baltimore*. Disponível em: <https://bit.ly/3aQXllk>

<sup>2</sup>Khan Academy. *Vírus de Animais e Humanos*. Disponível em: <https://bit.ly/2xXmr3n>



**Figura 2.** Mapa adaptado do Instituto Bloomberg, atualizado até 06 de abril de 2020, onde é possível visualizar a evolução dos casos globais de COVID-19.

dos Coronaviridae, os quais são popularmente denominados de coronavírus [3].

#### A. COVID-19

De acordo com [2], o coronavírus é um tipo de vírus que consiste de uma única cadeia de RNA, coberta por uma coroa de proteínas. Essa característica, a qual pode ser visualizada pela Figura 1, é responsável pelo nome dado a família do vírus. As espinhas provenientes dessa coroa permitem que o vírus possa se ligar ao hospedeiro e inserir seu material genético no núcleo da célula.

O coronavírus pode afetar mamíferos e aves, sendo que duas de suas cepas são responsáveis pela causa do resfriado comum em seres humanos. Entretanto, esse vírus também pode causar doenças respiratórias graves, dentre elas a Síndrome Respiratória Aguda Grave (SARS), descoberta em 2003, e a Síndrome Respiratória do Oriente Médio (MERS), descoberta em 2012<sup>3</sup>. Ambas as doenças emergiram de reservatórios de animais e foram responsáveis pela morte de um grande número de pessoas [4].

Os coronavírus que afetam os seres humanos eram considerados patógenos irrelevantes, pois normalmente causavam resfriados comuns em pessoas saudáveis. Porém, o surgi-

mento do SARS e do MERS provocou uma mudança neste posicionamento, devido ao fato de ambos serem responsáveis por causar epidemias globais com taxas alarmantes de morbidade e mortalidade.

Em dezembro de 2019, outra variação do coronavírus que afeta os seres humanos foi descoberta em Wuhan, capital da província de Hubei, na China [4]. Denominado de COVID-19 (Coronavirus Disease 2019), essa doença é causada pelo coronavírus da Síndrome Respiratória Aguda Grave 2 (SARS-CoV-2). Atualmente, o número de infectados pelo COVID-19 tem aumentado exponencialmente, conforme pode ser verificado pela Figura 2. Diante dessa rápida propagação e do aumento significativo no número de mortes em todo o planeta, a Organização Mundial de Saúde acabou por elevá-la ao status de pandemia<sup>4</sup>.

Para alguns, mesmo que o COVID-19 seja uma doença séria, a maioria das pessoas infectadas poderá desenvolver uma doença respiratória leve e moderada, a qual poderá ser curada sem maiores problemas. Entretanto, pessoas idosas e aquelas que apresentam algum problema médico pré-existente, tais como doenças cardiovasculares ou respiratórias, diabetes, câncer, entre outras, poderão desenvolver complicações mais graves, as quais podem evoluir para um quadro de pneumonia

<sup>3</sup>F. Forato. *Diferença entre COVID-19, SARS e MERS*. Disponível em: <https://bit.ly/2VlsIUn>

<sup>4</sup>World Healthy Organization. *General's Opening Remarks at the Media Briefing on COVID-19*. Disponível em: <https://bit.ly/3b0MgOE>

ou até mesmo levar a óbito.

Atualmente não existem vacinas ou medicamentos específicos para tratar a doença. Como forma de prevenção, para tentar diminuir a disseminação do vírus, as organizações de saúde recomendam boas práticas de higiene e orientam a quarentena.

### B. A IMPORTÂNCIA DA ANÁLISE DOS DADOS

Embora seja complicado alcançar uma previsão sobre o caminho percorrido por este surto viral, a análise dos dados obtidos até o momento pode contribuir para o desenvolvimento de estratégias a serem adotadas pelos órgãos responsáveis pela saúde pública, permitindo assim a implementação de medidas mais eficazes para o controle da expansão da epidemia [4]. Dessa maneira, inúmeras instituições têm disponibilizado dados relacionados ao COVID-19, permitindo assim que pesquisadores de diversas partes do mundo possam contribuir para a análise das informações.

Por meio da análise estatística é possível encontrar tendências e correlações relacionadas aos dados. Além disso, a análise das séries temporais presentes no conjunto de dados permite, dentre outras, coisas investigar o mecanismo gerador da série, realizar previsões de valores futuros, descrever o seu comportamento e encontrar periodicidades relevantes nos dados [5].

A conjectura por trás da análise de séries temporais é de que existe um sistema causal mais ou menos constante, relacionado com o tempo, o qual exerceu influência sobre os dados no passado. Esse sistema pode influenciar dados futuros e também é responsável por criar padrões não aleatórios que podem ser detectados em gráfico da série temporal ou por meio de outro processo estatístico. Logo, a análise de séries temporais permite identificar padrões aleatórios em uma variável de interesse. A observação desse comportamento permite realizar previsões sobre o futuro, orientando assim a tomada de decisões.

Além das técnicas clássicas fornecidas pela análise estatística básica, os dados de uma série temporal podem ser estudados mediante a aplicação de algoritmos de machine learning. Em parte, o aumento no uso desses tipos de técnicas foi impulsionado pelo resultado obtidos por empresas tais como Google, Facebook e Amazon, entre outras<sup>5</sup>.

Existem diversos modelos de machine learning que podem ser aplicados para a análise de séries temporais. Neste artigo, nosso objetivo é realizar a predição dos dados existentes em arquivos relacionados ao COVID-19. Como fonte de dados, utilizou-se um *dataset* disponibilizado pela plataforma Kaggle<sup>6</sup>.

A análise dos dados é realizada em conjunto com a aplicação do Teorema de Takens<sup>7</sup> para gerar os atributos preditivos e o atributo alvo. Esse teorema permite a reconstrução do

<sup>5</sup>C. E. Souza. *Séries Temporais com Machine Learning*. Disponível em: <https://bit.ly/2XivEhG>

<sup>6</sup>Kaggle. *COVID-19 Global Forecasting*. Disponível em: <https://bit.ly/2wmIOPz>

<sup>7</sup>Wikipedia. *Takens's Theorem*. Disponível em: <https://bit.ly/3aRKwYd>

espaço de fases por meio de séries temporais atrasadas de uma única observação do sistema original. De acordo com [6], o Teorema de Takens é uma ótima ferramenta para entendimento e estudo de sistemas dinâmicos na falta de informações completas sobre todas as variáveis do sistema, permitindo por meio da reconstrução, evidenciar informações antes não tangíveis somente a partir dos dados.

Para isso, utilizou-se técnicas de aprendizagem supervisada por meio da aplicação de modelos de regressão que utilizam conceitos relacionados às redes neurais artificiais do tipo Long Short Term Memory (LSTM) [7]. Como métrica de avaliação, verificou-se o valor para o erro médio quadrático, ou Mean Squared Error (MSE)<sup>8</sup>.

### II. DESCRIÇÃO DO CONJUNTO DE DADOS

PARA o estudo da pandemia causada pelo COVID-19, inúmeras organizações e entidades disponibilizaram conjuntos de dados, os quais encontram-se armazenados em vários repositórios online. Entretanto, muitos desses arquivos contêm apenas informações sobre o nome de um país ou região, a data em que uma determinada medição foi realizada e o número de casos confirmados e de fatalidades que foram registrados nessa data.

Dessa forma, para este artigo, além dos dados disponibilizados pela plataforma Kaggle, utilizou-se também informações de dados demográficos e socioeconômicos dos países<sup>9</sup>, <sup>10</sup>, <sup>11</sup> e <sup>12</sup>. Esses novos conjuntos de dados foram filtrados e as informações foram agregadas em um único arquivo, descrito a seguir.

O conjunto final de dados é composto por 22.270 observações e seu vetor de características contém ao todo 21 atributos. Esses atributos dizem respeito ao código ISO3 do país ou região onde a medição foi realizada, seu nome, sua província ou estado, os valores de sua latitude e longitude, seu PIB referente ao ano de 2019, o valor de sua população, a porcentagem de pessoas que vivem na área urbana, os valores de sua área territorial e de sua densidade demográfica, em km<sup>2</sup>, sua taxa de fertilidade, a média de idade de sua população, o percentual de fumantes existentes no país e a taxa de mortalidade causada por doenças pulmonares.

No arquivo, também foram acrescentadas informações sobre o total de leitos hospitalares para cada mil habitantes, os valores da temperatura média em graus Celsius e da umidade relativa do ar, registrados entre os meses de janeiro e março de 2020, e a data da ocorrência do primeiro caso de COVID-19 no local.

<sup>8</sup>J. J. MAE and RMSE - Which Metric is Better? Disponível em: <https://bit.ly/2XWtliO>

<sup>9</sup>Johns Hopkins. CSSE COVID-19 Dataset. Disponível em: <https://bit.ly/2V90R5b>

<sup>10</sup>Worldometer. Countries in the World by Population (2020) Forecasting. Disponível em: <https://bit.ly/2yhE3al>

<sup>11</sup>Kaggle. COVID-19: Additional Statistics Forecasting. Disponível em: <https://bit.ly/2RC7P0u>

<sup>12</sup>Kaggle. COVID-19 Enriched Dataset Week 2. Disponível em: <https://bit.ly/2wIsuJ8>

**Tabela 1.** Vetor de Característica do Conjunto Final de Dados

	Atributo	Tipo	Escala	Exemplo
1	Código do País	Qualitativo	Nominal	BRA
2	País/Região	Qualitativo	Nominal	Brazil
3	Província/Estado	Qualitativo	Nominal	
4	Latitude	Quant. contínuo	Racional	-14.235
5	Longitude	Quant. contínuo	Racional	-51.9253
6	PIB 2019	Quant. contínuo	Racional	1847020.0
7	População	Quant. contínuo	Racional	212559417.0
8	População Urbana	Quant. contínuo	Intervalar	88.8
9	Área (Km <sup>2</sup> )	Quant. contínuo	Racional	8358140.0
10	Densidade (Km <sup>2</sup> )	Quant. contínuo	Racional	25.0
11	Taxa de Fertilidade	Quant. contínuo	Racional	1.7
12	Média de Idade	Quant. contínuo	Racional	33.0
13	Taxa de Fumantes	Quant. contínuo	Intervalar	15.3
14	Taxa de Mort. Pulmonar	Quant. contínuo	Intervalar	26.57
15	Total de Leitos	Quant. contínuo	Racional	2.2
16	Temperatura Média	Quant. contínuo	Intervalar	25.82
17	Taxa de Umidade	Quant. contínuo	Intervalar	81.69
18	Data do Primeiro Caso	Quant. discreto	Intervalar	2/24/2020
19	Data	Quant. discreto	Intervalar	1/22/2020
20	Casos Confirmados	Quant. discreto	Racional	0
21	Fatalidades	Quant. discreto	Racional	0

O valor do atributo *PIB 2019* está representado em milhões de dólares; *Taxa de Mort. Pulmonar* = Taxa de Mortalidade por Doenças Pulmonares; *Total de Leitos* = Total de Leitos Hospitalares, para cada mil habitantes; Os valores da *Temperatura Média* e *Taxa de Umidade* foram coletados entre os meses de janeiro e março de 2020.

A coluna *Exemplo* exibe uma das possíveis entradas para os atributos. Observe que, para o caso do Brasil, não temos os valores para o atributo *Província/Estado*.

Por último, adicionou-se os atributos referentes à data diária da medição dos casos de COVID-19, correspondentes ao período entre 01 de janeiro de 2020 até 15 de abril de 2020, junto com o número de casos que foram confirmados e o número de fatalidades que ocorreram nessa data. A Tabela 1 exibe o vetor de características para esse conjunto final de dados, com a descrição de seus atributos.

A próxima seção descreve uma visão global para esse conjunto de dados, obtida por meio da sintetização de suas informações.

### III. ANÁLISE EXPLORATÓRIA DOS DADOS

PARA que se possa começar a etapa de pré-processamento, é importante verificar as principais características do conjunto de dados. Dessa forma, no início deste trabalho, realizou-se o procedimento denominado Análise Exploratória de Dados (AED). A AED é uma ferramenta muito importante, pois permite obter um resumo das principais características existentes em um conjunto de dados. O resultado obtido orienta aquilo que deve ser feito na seção de pré-processamento.

O conjunto de dados utilizado para este trabalho é composto por 21 atributos e 22.270 registros. Esses registros correspondem às observações realizadas entre o período de

**Tabela 2.** Atributos, Tipos de Dados, Valores Não-Nulos e Nulos

	Atributo	Tipo de Dados	Não Nulos	Nulos
1	Código do País	object	20.995	1.275
2	País/Região	object	22.270	0
3	Província/Estado	object	6.800	15.470
4	Latitude	float	22.270	0
5	Longitude	float	22.270	0
6	PIB 2019	float	20.995	1.275
7	População	float	21.250	1.020
8	População Urbana	float	20.825	1.445
9	Área (Km <sup>2</sup> )	float	21.250	1.020
10	Densidade (Km <sup>2</sup> )	float	21.250	1.020
11	Taxa de Fertilidade	float	20.740	1.530
12	Média de Idade	float	20.740	1.530
13	Taxa de Fumantes	float	16.915	5.355
14	Taxa de Mort. Pulmonar	float	20.315	1.955
15	Total de Leitos	float	20.995	1.275
16	Temperatura Média	float	18.615	3.655
17	Taxa de Umidade	float	18.615	3.655
18	Data do Primeiro Caso	object	20.995	1.275
19	Data	object	22.270	0
20	Casos Confirmados	int	22.270	0
21	Fatalidades	int	22.270	0

01 de janeiro de 2020 até 15 de abril de 2020, cujo vetor de características pode ser consultado na Tabela 1.

A Tabela 2 exibe o tipo de dados para cada um desses atributos, juntamente com o total de valores não-nulos e nulos, que estão armazenados no arquivo.

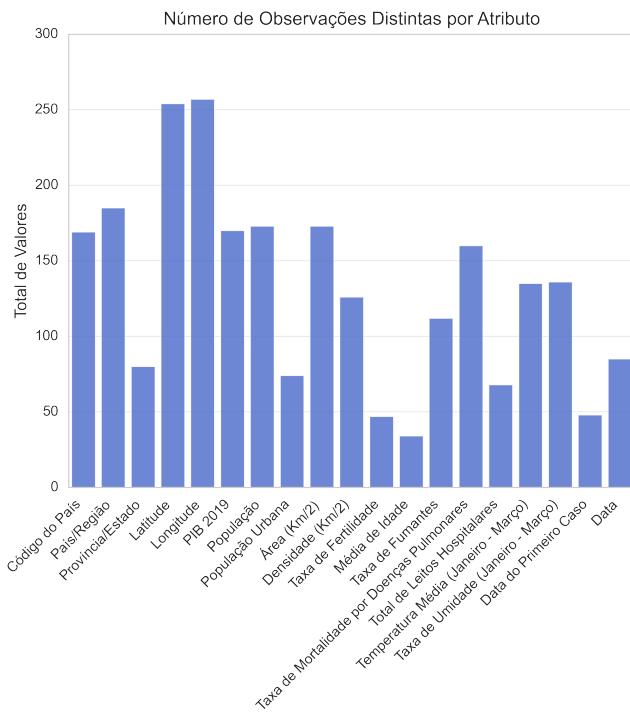
O tipo de dados *object* deve ser convertido para o formato adequado, durante a etapa de pré-processamento dos dados. Dependendo do objetivo, linhas onde o valor do atributo seja nulo podem ser removidas ou agrupadas, de acordo com os atributos Código do País, País/Região, Província/Estado, dentre outros. Além disso, novos dados também podem ser imputados para preencher os valores ausentes.

Os valores ausentes referem-se aos campos vazios ou sem valores atribuídos a eles. Essa situação geralmente ocorre devido a erros durante a entrada de dados, falhas nos processos de coleta ou em operações utilizadas para unir várias colunas de tabelas ou arquivos diferentes.

O número de valores distintos para cada atributo, com exceção de Casos Confirmados e Fatalidades, pode ser visualizado no gráfico da Figura 3. Para esses dois atributos, o total de valores distintos corresponde aos números 2.335 e 652, respectivamente.

O valor de 22.270 registros armazenados no arquivo de dados é referente a uma combinação das observações capturadas para a data de leitura dos casos confirmados, levando-se em consideração um país ou sua província. Dessa forma, para cada uma das 85 datas distintas, existem 262 medições que foram cadastradas.

Observando a Tabela 2, pode-se notar que o atributo Província/Estado possui muitos valores nulos, o que não ocorre com o atributo País/Região. Isso significa que nem todos os



**Figura 3.** Número de observações distintas para cada atributo do conjunto de dados.

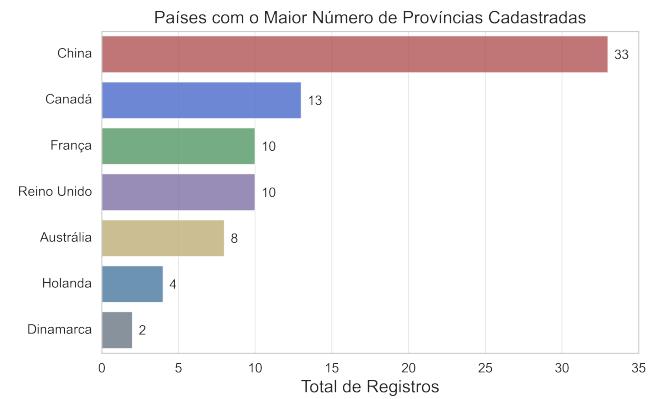
países tiveram suas províncias ou estados cadastrados separadamente. Por meio de uma consulta no arquivo de dados, foi possível perceber que para o país onde o cadastro não foi realizado pela província ou estado, existem no máximo 85 registros armazenados pela data distinta da medição. Em outras situações, esse número pode variar. Um exemplo é a China, onde o número de medições para o mesmo período é de 2.805.

Analizando a situação, foi possível verificar que isso ocorreu porque o arquivo de dados contém o registro de 33 províncias chinesas. Observando os valores para atributo País/Região, em relação ao atributo Província/Estado, foi constatado que as medições realizadas em 178 países não levaram em consideração suas províncias ou estados. Esse é o caso, por exemplo, do Brasil. Durante a etapa de pré-processamento, essas ocorrências devem ser levadas em consideração e tratadas conforme os objetivos desejados, para que os resultados se mantenham coerentes. A Figura 4 exibe um comparativo entre os países com o maior número de províncias ou estados armazenados no conjunto de dados.

#### A. ANÁLISE DE DISTRIBUIÇÃO DAS VARIÁVEIS

Para verificar a distribuição de alguns dos atributos armazenados no conjunto de dados utilizou-se um gráfico de densidade. Esse tipo de representação permite o estudo da distribuição de variáveis numéricas, o que auxilia na descoberta de possíveis erros ou anomalias.

Analizando o resultado obtido, apresentado na Figura 6, é possível notar que a maioria da população urbana está



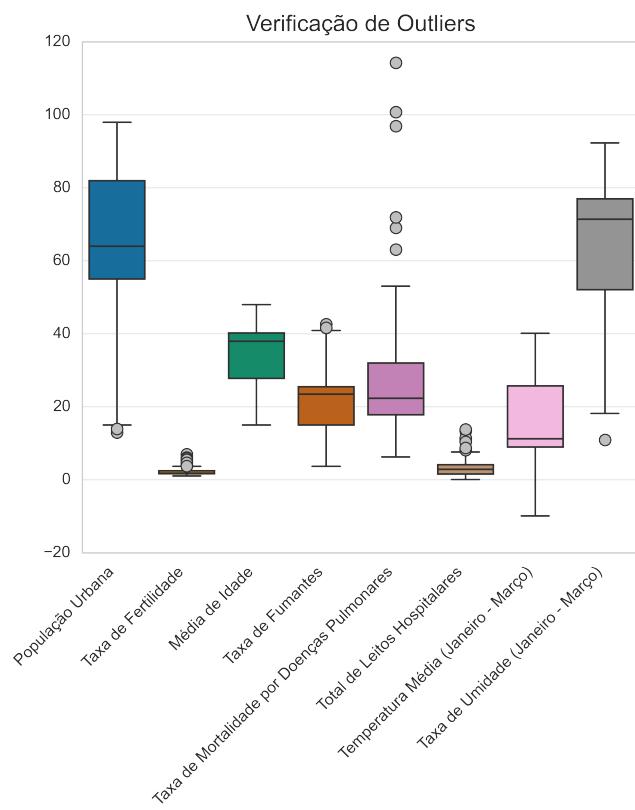
**Figura 4.** Gráfico com os países que possuem o maior número de províncias cadastradas no arquivo de dados. Para os demais, os registros dos casos não foram armazenados separadamente para cada província.

concentrada no valores de 60% e 80%. A taxa de fertilidade informa que a maior parte das mulheres possui por volta de 2 filhos. A média de idade da população está situada na faixa dos 40 anos. A taxa de fumantes indica uma maior concentração em torno de 25% da população. Para a taxa de mortalidade por doenças pulmonares, a cada 200 mil pessoas, o gráfico indica uma maior concentração em torno do valor 20. O total de leitos hospitalares para cada 1000 pessoas possui uma ligeira representatividade nos valores 3 e 4. Por último, vemos que a temperatura média entre os meses de janeiro e março oscilou em torno de 10 graus Celsius e a taxa de umidade para o mesmo período teve picos variando entre 40% e 75%.

Continuando a análise, verificou-se se os dados possuem valores que fogem um pouco da normalidade, os quais também são conhecidos por *outliers*. Dependendo dos objetivos e dos algoritmos utilizados, esses valores podem interferir no resultado. Para visualizar essa situação, utilizou-se um *boxplot*, representado na Figura 5.

De acordo com o resultado, pode-se observar a presença de *outliers* nos atributos População Urbana, Taxa de Fertilidade, Taxa de Fumantes, Taxa de Mortalidade por Doenças Pulmonares e Taxa de Umidade. Alguns desses valores demonstram certa coerência, pois correspondem às características de alguns dos países cadastrados no conjunto de dados. Por exemplo, temos países onde o percentual da população urbana é baixo. Além disso, existem situações onde o número de filhos por cada mulher, o número de fumantes, o número de mortes causada por doenças pulmonares ou o total de leitos hospitalares pode ser considerado demasiadamente alto. Essa discrepância também pode ser verificada pela análise da distribuição das variáveis apresentada na Figura 6.

Verificando os outros atributos armazenados no conjunto de dados, notou-se que os números dos casos confirmados de COVID-19 e das fatalidades apresentam os maiores *outliers*. O desvio padrão para esses dois atributos é muito alto, sendo representado pelos números 14556.72 e 890.46, respectivamente, o que indica que os pontos estão espalhados por uma



**Figura 5.** Boxplot para algumas das variáveis existentes no conjunto de dados, onde é possível verificar a presença de *outliers*.

ampla gama de valores.

Para analisar a relação entre os atributos do conjunto de dados gerou-se a matriz de correlação, a qual pode ser visualizada na Figura 7.

O coeficiente de correlação mede a direção e o grau de associação linear entre as variáveis. Valores mais próximos de -1 ou +1 indicam uma forte associação entre as variáveis, enquanto valores próximos de zero sinalizam o contrário. Um sinal positivo indica que o valor de uma variável aumenta quando a outra também aumenta. O sinal negativo indica o decrescimento de uma variável em relação a outra. Observando os valores dos coeficientes de correlação, pode-se entender qual atributo entre os pares é mais significativo para a construção dos modelos desejados.

Analizando os valores indicados na Figura 7, pode-se observar uma forte correlação positiva entre os atributos que armazenam os dados do PIB, da população e da área de um país, bem como do número de casos confirmados e de suas fatalidades. Percebeu-se também, que a média da idade interfere na taxa de fertilidade das mulheres. Além disso, outro fato interessante é que a taxa de fumantes não tem muita correlação com a taxa de mortalidade por doenças pulmonares.

Por meio da análise exploratória, pode-se compreender como o nosso conjunto de dados foi armazenado. A próxima seção descreve o pré-processamento das informações, vi-

sando preparar nossos dados para a aplicação dos algoritmos de aprendizagem de máquina.

#### IV. PRÉ-PROCESSAMENTO DOS DADOS

**A** PÓS o término da etapa de análise exploratória dos dados é possível compreender como nossos dados estão organizados. Com isso, em algumas situações, deve-se realizar o pré-processamento das informações, com o objetivo de organizar melhor nossa base de dados.

A etapa de pré-processamento dos dados é composta por um conjunto de atividades que envolvem preparação, organização e estruturação dos dados. Trata-se de uma etapa fundamental que precede a realização de análises e previsões.

Entre alguns dos principais problemas que podemos encontrar dentro de um conjunto de dados temos a falta dos valores para alguns de seus atributos, a presença de *outliers*, a utilização de escalas diferentes para os mesmos valores, entre outros. Esses problemas geralmente são identificados na etapa de análise exploratória. Na etapa de pré-processamento, pode-se utilizar diferentes técnicas [9], tais como agregação, amostragem, redução de dimensionalidade, seleção de subconjuntos de recursos, transformação de variáveis, etc., para corrigir essas anomalias. Essas técnicas costumam ser empregadas dentro de etapas que envolvem a limpeza, transformação e redução dos dados<sup>13</sup>.

Em seguida, apresenta-se uma descrição dos passos que foram utilizados durante a etapa de pré-processamento de dados realizada para este trabalho.

##### A. LIMPEZA DOS DADOS

O conjunto de dados original possui um total de 22.270 observações, cujo vetor de características contém 21 atributos. Analisando os valores armazenados, notou-se que 170 observações eram correspondentes aos dados coletados em navios de cruzeiro<sup>14</sup>, os quais não estavam relacionados a nenhum país. Para essas observações, não havia sentido no armazenamento de diversos atributos. Dessa forma, optou-se pela remoção dessas linhas.

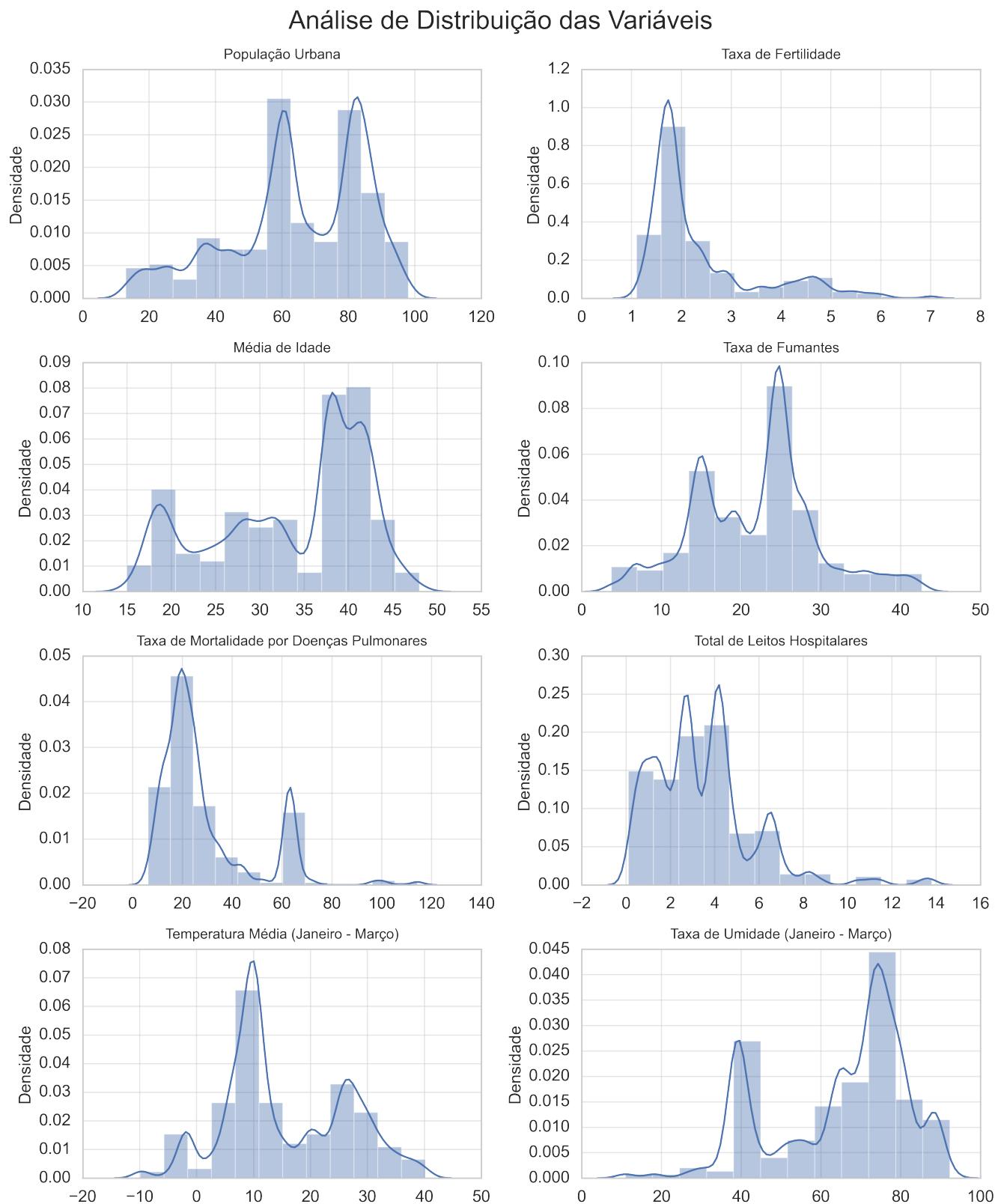
Criou-se uma listagem contendo o nome de 30 navios de cruzeiro, tais como Diamond Princess, MS Zaandam, MS Westerdam, entre outros, e realizou-se a remoção das observações presentes no arquivo de dados que estavam relacionadas ao nome de algum desses navios. Ao término desse procedimento, o arquivo de dados passou a conter somente observações referentes aos países onde os casos de COVID-19 foram registrados.

##### B. IMPUTAÇÃO DE NOVOS VALORES

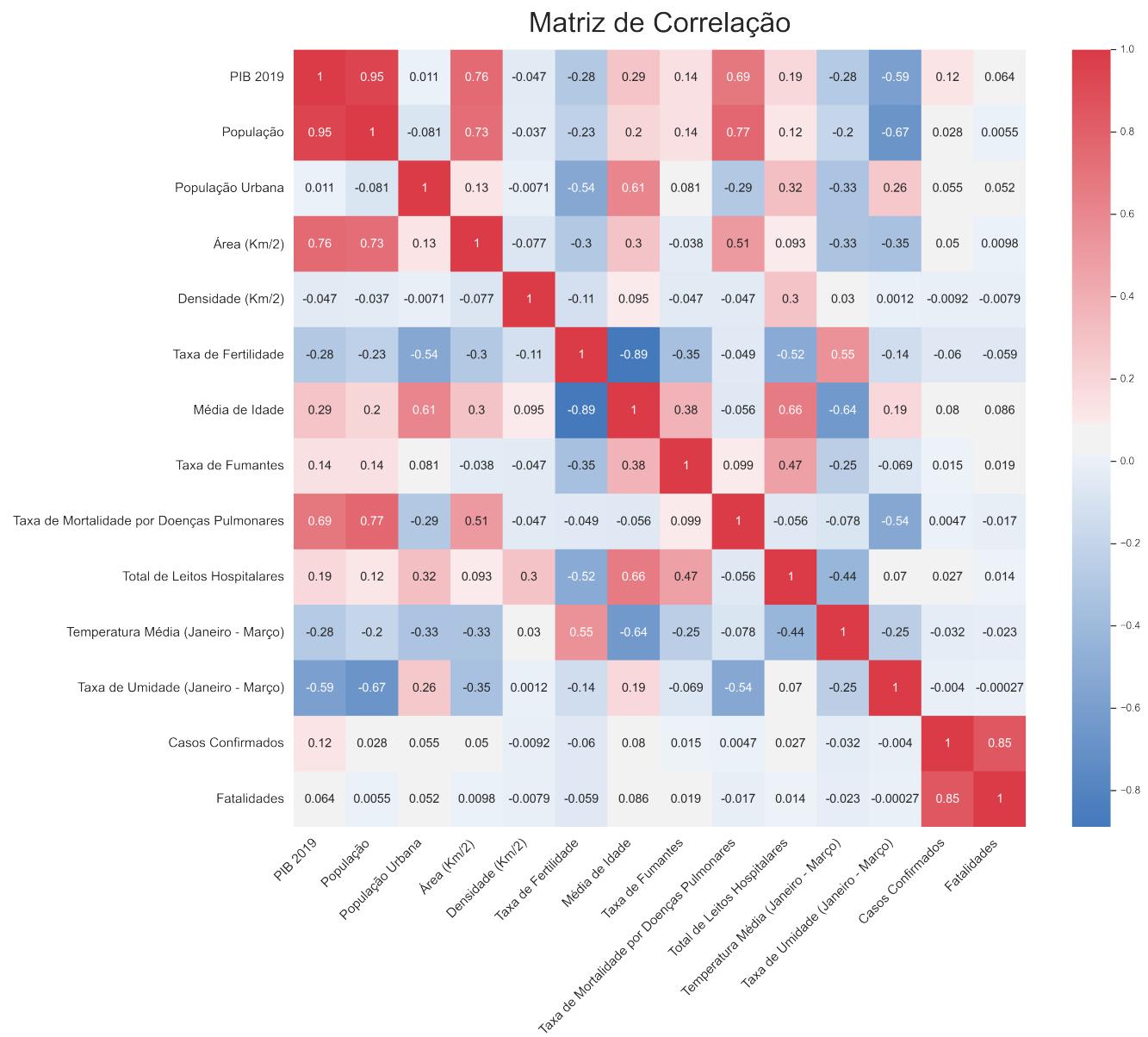
Após a remoção das observações que não foram atribuídas a um determinado país, realizou-se a imputação de novos valores, com o objetivo de substituir aqueles que não haviam sido fornecidos. Ao analisar o vetor de características,

<sup>13</sup>Pedro César Tebaldi Gomes. *Conheça as Etapas do Pré-Processamento de dados*. Disponível em: <https://bit.ly/2WxVvjB>

<sup>14</sup>Wikipedia. *Pandemia de COVID-19 em navios de cruzeiro*. Disponível em: <https://bit.ly/2yvhgIM>



**Figura 6.** Gráfico com a análise de distribuição para algumas das variáveis existentes no conjunto de dados, onde é possível notar a maior concentração dos valores.



**Figura 7.** Matriz de correlação para os atributos numéricos do conjunto de dados, onde é possível visualizar a força da associação entre algumas variáveis.

observou-se que diversos atributos relacionados aos dados demográficos ou socioeconômicos estavam ausentes. Para a maioria deles, realizou-se uma busca em *sites* específicos, afim de encontrar seus valores.

Para transformar nosso conjunto de dados em um arquivo do tipo atributo-valor, removeu-se o atributo País/Região e substituiu-se o código ISO3 de cada localidade por um valor numérico. Esses novos valores correspondem ao atributo Código do País. Os valores ausentes para o código ISO3 foram obtidos após a consulta de fontes específicas<sup>15</sup>.

Para imputar os valores ausentes referentes aos atribu-

<sup>15</sup>IBAN. *Country Codes Alpha-2 and Alpha-3*. Disponível em: <https://bit.ly/3cjkptG>

tos PIB 2019, População, População Urbana, Área (Km²), Densidade (Km²), Taxa de Fertilidade e Média de Idade, utilizou-se as informações disponibilizadas em Country Economy<sup>16</sup>, Worldometer<sup>17</sup> e Index Mundi<sup>18</sup>. Além desses, também foram consultadas informações disponibilizadas pela Central Intelligence Agency<sup>19</sup>.

Em relação aos valores ausentes para o atributo Taxa de

<sup>16</sup>Country Economy. *Dados Econômicos e Demográficos*. Disponível em: <https://bit.ly/2YDvToc>

<sup>17</sup>Worldometer. *Population of Countries*. Disponível em: <https://bit.ly/3ccId2>

<sup>18</sup>Index Mundi. *Countries*. Disponível em: <https://bit.ly/2W4m5lm>

<sup>19</sup>Central Intelligence Agency. *The World Factbook*. Disponível em: <https://bit.ly/3fqVNkL>

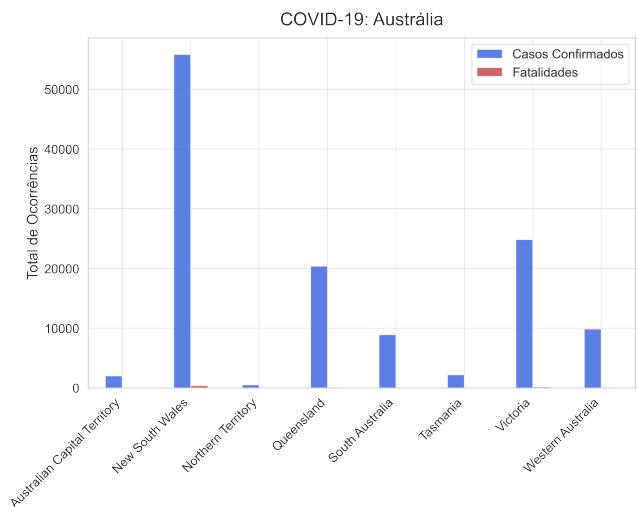
Fumantes, verificou-se as informações disponíveis em The Tobacco Atlas<sup>20</sup> e Nation Master<sup>21</sup>. Entretanto, os valores para os países Liechtenstein, Monaco, Western Sahara e Taiwan não foram encontrados. Dessa forma, optou-se por atualizar essas observações imputando o valor da média obtida para o atributo Taxa de Fumantes. Realizou-se um procedimento semelhante para atualizar os valores ausentes para o atributo Total de Mortalidades por Doenças Pulmonares, onde se consultou os dados disponíveis em World Health Rankings<sup>22</sup>. Segundo o site, os dados para os países Andorra, Cook Islands, Dominica, Marshall Islands, Monaco, Nauru, Niue, Palau, Saint Kitts, San Marino e Tuvalu foram excluídos pela Organização Mundial de Saúde. Assim, para esses países, optou-se por imputar o valor zero para o atributo em questão. Consultou-se o The World Bank<sup>23</sup> para obter os valores que estavam faltando para o atributo Total de Leitos Hospitalares. Para imputar os valores da Temperatura e da Umidade, acessou-se o Weather and Climate<sup>24</sup>.

Apesar do nosso conjunto de dados ter um total de 22.100 observações, obtidas como resultado da remoção das entradas referentes aos navios de cruzeiro, cada país possui 85 medições que foram coletadas entre janeiro e abril de 2020. No conjunto de dados, existem informações de 169 países diferentes cadastrados, o que totaliza 14.365 observações. O restante, correspondente a 7.735 observações, diz respeito aos países cuja entrada foi cadastrada por província ou estado, como é o caso da China, Canadá, França, Reino Unido, Austrália, Holanda e Dinamarca. Para os demais países, como o Brasil por exemplo, a observação não foi armazenada dessa forma. A Figura 8, por exemplo, exibe o número de casos confirmados e de fatalidades registradas no período, para os territórios da Austrália.

Como o objetivo deste trabalho é realizar a predição para o número de casos confirmados de COVID-19, optou-se por remover os atributos Província/Estado, Data do Primeiro Caso e Fatalidades. Além disso, os registros pertencentes a cada Província/Estado foram agrupados em seus respectivos países. Por último, o atributo Data foi utilizado para indexar o novo conjunto de dados. O arquivo final ficou com um total de 15.555 observações e 17 atributos.

A Tabela 3 exibe o total de valores nulos antes e depois da etapa de pré-processamento de dados, juntamente com os atributos que foram selecionados para a etapa de predição.

Após a etapa de pré-processamento, o conjunto de dados foi exportado para um arquivo do tipo CSV, denominado *covid19-atributo-valor.csv*. Os dados desse arquivo foram



**Figura 8.** Total de casos confirmados e de fatalidades, registrados para os territórios australianos, durante os meses de janeiro e abril de 2020.

**Tabela 3.** Total de Valores Nulos Antes e Depois do Pré-processamento

	Atributo	Nulos - Antes	Nulos - Depois
1	Data	0	0
2	Código do País	1.275	0
3	Latitude	0	0
4	Longitude	0	0
5	PIB 2019	1.275	0
6	População	1.020	0
7	População Urbana	1.445	0
8	Área (Km <sup>2</sup> )	1.020	0
9	Densidade (Km <sup>2</sup> )	1.020	0
10	Taxa de Fertilidade	1.530	0
11	Média de Idade	1.530	0
12	Taxa de Fumantes	5.355	0
13	Taxa de Mort. Pulmonar	1.955	0
14	Total de Leitos	1.275	0
15	Temperatura Média	3.655	0
16	Taxa de Umidade	3.655	0
17	Casos Confirmados	0	0

utilizados para a aplicação dos algoritmos de aprendizado de máquina.

A próxima seção apresenta a descrição para o algoritmo de aprendizado de máquina utilizado neste trabalho.

## V. DESCRIÇÃO DO MODELO DE PREDIÇÃO

**A** PÓS a etapa de pré-processamento, pode-se aplicar algoritmos de aprendizado de máquina para descobrir novas informações relacionadas ao nosso conjunto de dados. Alguns desses algoritmos envolvem tarefas de regressão, onde geralmente se procura por valores os quais não é possível estimar inicialmente. Para este artigo, pretende-se estudar a evolução do número de casos confirmados de COVID-19.

A análise de regressão consiste em técnicas para modelar o relacionamento entre uma variável dependente e uma ou

<sup>20</sup>The Tobacco Atlas. *The Tobacco Atlas*. Disponível em: <https://bit.ly/3b93WXP>

<sup>21</sup>Nation Master. *Total Adult Smokers: Countries Compared*. Disponível em: <https://bit.ly/2W5nesH>

<sup>22</sup>World Health Rankings. *Lung Disease*. Disponível em: <https://bit.ly/2SDzDSM>

<sup>23</sup>The World Bank. *Hospital Beds*. Disponível em: <https://bit.ly/2A2PyDH>

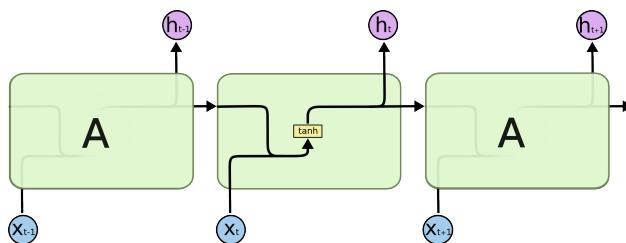
<sup>24</sup>Weather and Climate. *How's the Weather in Other Parts of the World?*. Disponível em: <https://bit.ly/2YGwSDV>

mais variáveis independentes, as quais também costumam ser conhecidas como variáveis explicativas ou preditores [10]. Existem diferentes tipos de regressão. Para este artigo, utilizou-se as redes neurais do tipo LSTM.

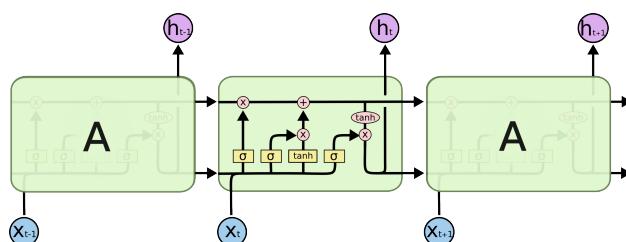
Os problemas que envolvem a previsão de séries temporais constituem um tipo difícil de problema de modelagem preditiva, pois as séries temporais também adicionam a complexidade de uma dependência de sequência entre as variáveis de entrada<sup>25</sup>. Embora existam várias metodologias clássicas que podem ser utilizadas para auxiliar nessa tarefa, os modelos que utilizam redes neurais do tipo LSTM se tornaram muito úteis para se lidar com esses tipos de dados<sup>26</sup>.

As redes neurais LSTM são um tipo especial de rede neural recorrente (RNN), capaz de aprender dependências de longo prazo. Elas foram introduzidas por Hochreiter e Schmidhuber em 1997 e foram refinadas e popularizadas em diversos estudos<sup>27</sup>.

Todas as redes neurais recorrentes têm a forma de uma cadeia de módulos repetitivos da rede neural, como ilustra a Figura 9. Uma rede neural LSTM também possui essa estrutura em cadeia. Entretanto, o módulo de repetição apresenta uma configuração diferente, como demonstra a Figura 10.

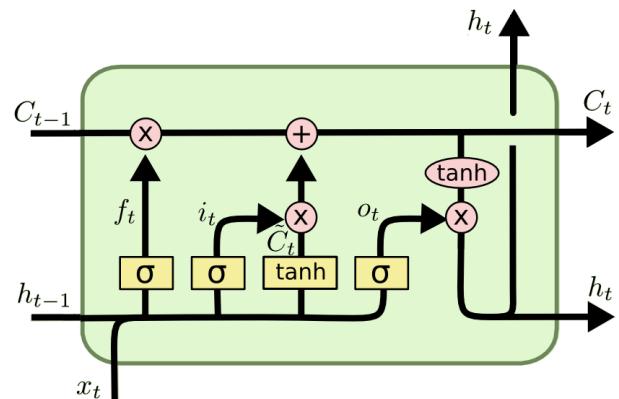


**Figura 9.** Módulo de repetição em uma rede neural recorrente padrão, contendo uma única camada.



**Figura 10.** Módulo de repetição em uma rede LSTM, contendo quatro camadas de interação.

Essa configuração apresenta uma estrutura em cadeia, que contém quatro redes neurais e diferentes blocos de memória chamados células. A informação é retida pelas células e as manipulações de memória são feitas por portões, ou *gates*. A Figura 11 ilustra o caminho percorrido dentro de uma célula.



**Figura 11.** Caminho percorrido dentro de um módulo de uma rede LSTM.

A passagem por uma célula de uma rede neural LSTM são dadas pelo conjunto de equações listadas em (1):

$$\begin{aligned}
f_t &= \sigma(x_t * U_f + h_{t-1} * W_f) \\
\tilde{C}_t &= \tanh(x_t * U_c + h_{t-1} * W_c) \\
i_t &= \sigma(x_t * U_i + h_{t-1} * W_i) \\
o_t &= \sigma(x_t * U_o + h_{t-1} * W_o) \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\
h_t &= o_t * \tanh(C_t)
\end{aligned} \tag{1}$$

Onde,

- $x_t$ : vetor de entrada;
  - $h_{t-1}$ : saída da célula anterior;
  - $C_{t-1}$ : memória da célula anterior;
  - $h_t$ : saída da célula atual;
  - $C_t$ : memória da célula atual;
  - $W, U$ : vetores de peso para os portões *forget* ( $f$ ), *candidate* ( $c$ ), *input* ( $i$ ) e *output* ( $o$ ).

Neste artigo, utilizou-se estes algoritmos em combinação com a seleção dos *features* e do atributo alvo baseados nos conceitos definidos pelo Teorema de Takens [11]. Esse teorema fornece a base teórica para a análise de séries temporais geradas a partir de sistemas determinísticos não lineares. Segundo [12], ele permite reconstruir a dinâmica do sistema analisado e desconhecido, a partir de uma série temporal observada por meio da elaboração de um novo espaço de estados, recuperando a possível estrutura geométrica imersa nesse espaço.

Considerando uma série temporal  $\{x(t_1), x(t_2), \dots, x(t_n)\}$ , sendo  $n$  o tamanho da série temporal, o espaço de estados formados pelas coordenadas de tempo de atraso pode ser escrito como vetores  $\xi_i$ , os quais representam a posição do ponto no espaço de imersão [13], como demonstra as equações em (2), onde  $i$  varia de 1 até  $n$ ,  $(t_{i+1} - t_i) = \Delta t$ , para  $\tau = \Delta t$ ,  $m$  está relacionado a dimensão requerida e  $\tau$  é o tempo de atraso. A escolha adequada de  $\tau$  é importante para a reconstrução do espaço de estados.

<sup>25</sup>J. Brownlee. *Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras*. Disponível em: <https://bit.ly/2AuRwNo>

**26** V. Valkov. *Time Series Forecasting with LSTMs for Daily Coronavirus Cases using PyTorch in Python*. Disponível em: <https://bit.ly/3dSseat>

*Cases using PyTorch in Python.* Disponível em: <https://bit.ly/3dSseal>  
*27C. Olah. Understanding LSTM Networks.* Disponível em: <https://bit.ly/2AzGeaw>

$$\begin{aligned}
 \xi_1 &= (x(t_1), x(t_1 + \tau), x(t_1 + 2\tau), \dots, x(t_1 + (m-1)\tau)), \\
 \xi_2 &= (x(t_2), x(t_2 + \tau), x(t_2 + 2\tau), \dots, x(t_2 + (m-1)\tau)), \\
 &\vdots \\
 \xi_i &= (x(t_i), x(t_i + \tau), x(t_i + 2\tau), \dots, x(t_i + (m-1)\tau)), \\
 &\vdots
 \end{aligned} \tag{2}$$

A próxima seção apresenta o resultado obtido durante a aplicação da rede neural LSTM para a tarefa de predição dos casos confirmados de COVID-19.

## VI. PREDIÇÃO DOS CASOS UTILIZANDO REDES LSTM

**P**ARA a tarefa de predição dos casos confirmados de COVID-19, optou-se por variar o número de atrasos onde cada medição foi realizada, com o objetivo de verificar a existência de alguma variação entre elas.

Dessa forma, utilizou-se os valores 1, 3, 7, 15, 30 e 60, referentes ao  $\Delta t$  indicativo para esses números de atrasos. Além disso, definiu-se três conjuntos de valores para os hiperparâmetros utilizados pela rede neural LSTM.

Para a aplicação da rede neural, escolheu-se por variar o número de neurônios de entrada e o número de épocas no qual o algoritmo de aprendizado deveria atuar em cada conjunto de dados de treinamento. Variou-se também o tamanho de lote, ou *batch size*, utilizado para o processamento do número de amostras que deveriam ser trabalhadas antes da atualização dos parâmetros internos do modelo. Por último, adotou-se um valor de 20% para a camada de prevenção ao *overfitting*. Como algoritmo de otimização do modelo utilizou-se o Adam<sup>28</sup>. Para a estimativa dos erros, utilizou-se o MSE<sup>29</sup>. A Tabela 4 exibe as configurações utilizadas para os hiperparâmetros da rede neural.

**Tabela 4.** Configurações Utilizadas para a Rede Neural LSTM

	Parâmetro	Configuração 1	Configuração 2	Configuração 3
1	Neurônios	4	20	50
2	Épocas	10	100	500
3	Batch Size	10	50	72
4	Dropout	0.2	0.2	0.2
5	Otimizador	Adam	Adam	Adam
6	Estimador de Erro	MSE	MSE	MSE

Para a seleção dos atributos utilizados pelo modelo de predição desconsiderou-se aqueles com a informação geográfica dos países (Código do País, Latitude e Longitude). Em seguida, aplicou-se o método Wrapper<sup>30</sup>, onde se considerou como alvo o atributo Casos Confirmados.

<sup>28</sup>J. Brownlee. *Gentle Introduction to the Adam Optimization Algorithm for Deep Learning*. Disponível em: <https://bit.ly/3e3cgdM>

<sup>29</sup>J. Brownlee. *Time Series Forecasting Performance Measures With Python*. Disponível em: <https://bit.ly/2AqAkZR>

<sup>30</sup>A. Shetye. *Feature Selection with sklearn and Pandas*. Disponível em: <https://bit.ly/3cVeHhm>

A aplicação desse método retornou sete atributos, os quais foram utilizados na etapa de predição: PIB 2019, População, Área (KM/2), Taxa de Fertilidade, Total de Leitos Hospitalares e Temperatura Média (Janeiro - Março). Além disso, removeu-se as instâncias onde o número de casos confirmados era igual a zero.

Após a configuração dos valores para os hiperparâmetros da rede neural LSTM e seleção dos atributos, executou-se o código para a criação do modelo de predição, onde se utilizou 80% dos dados para a etapa de treinamento e 20% para a etapa de testes (método de validação por Holdout [14]). Por último, para os conjuntos de predição obtidos, referentes aos valores de cada  $\Delta t$  utilizado, calculou-se também suas respectivas mediana, média e desvio padrão. O resultado obtido para essas métricas pode ser visualizado nas Tabelas 5, 6 e 7.

**Tabela 5.** Medidas Estatísticas Obtidas para a Configuração 1

$\Delta t$	MSE	Mediana	Média	Desvio Padrão
1	7.48e+12	2.85e+07	2.85e+07	9.62e+04
3	7.39e+12	2.86e+07	2.86e+07	1.36e+05
7	4.72e+12	2.91e+07	2.9e+07	1.95e+05
15	2.94e+12	2.95e+07	2.95e+07	2.91e+05
30	6.82e+12	2.87e+07	2.87e+07	8.01e+04
60	3.16e+12	2.95e+07	2.95e+07	2.61e+05

**Tabela 6.** Medidas Estatísticas Obtidas para a Configuração 2

$\Delta t$	MSE	Mediana	Média	Desvio Padrão
1	6.16e+11	3.04e+07	3.03e+07	6.76e+05
3	9.45e+11	3.01e+07	3.01e+07	7.34e+05
7	9.29e+11	3.01e+07	3.01e+07	7.31e+05
15	4.37e+11	3.05e+07	3.05e+07	7.09e+05
30	4.76e+11	3.04e+07	3.04e+07	8.34e+05
60	1.08e+12	3.01e+07	3.01e+07	6.22e+05

**Tabela 7.** Medidas Estatísticas Obtidas para a Configuração 3

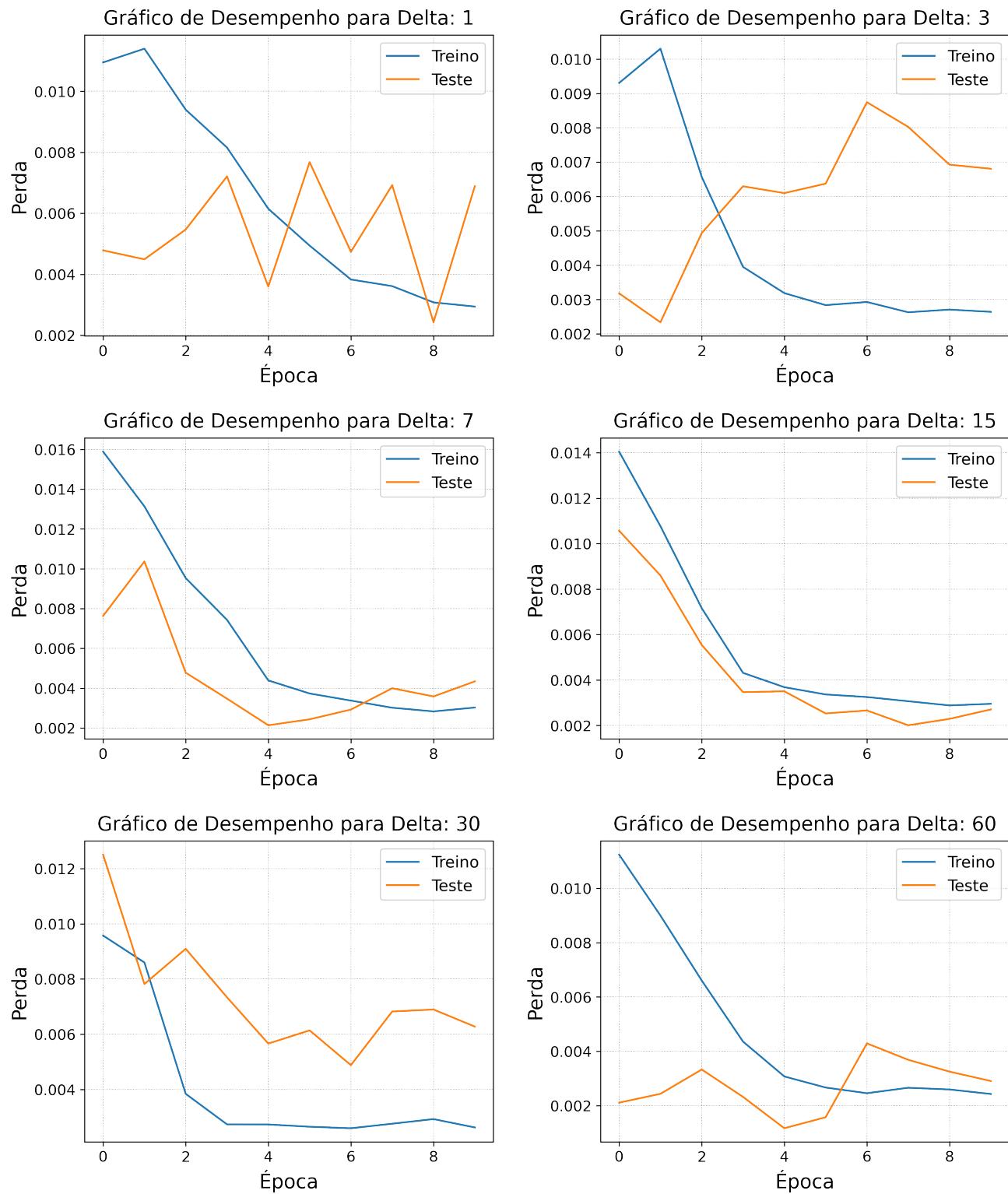
$\Delta t$	MSE	Mediana	Média	Desvio Padrão
1	1.1e+11	3.08e+07	3.08e+07	8.97e+05
3	8.14e+11	3.02e+07	3.02e+07	6.87e+05
7	2.05e+10	3.11e+07	3.11e+07	1.02e+06
15	3.95e+10	3.11e+07	3.12e+07	1.23e+06
30	7.26e+10	3.07e+07	3.07e+07	1.15e+06
60	4.7e+11	3.04e+07	3.04e+07	7.85e+05

A próxima seção apresenta a discussão a respeito dos resultados obtidos.

## VII. VALIDAÇÃO E DISCUSSÃO

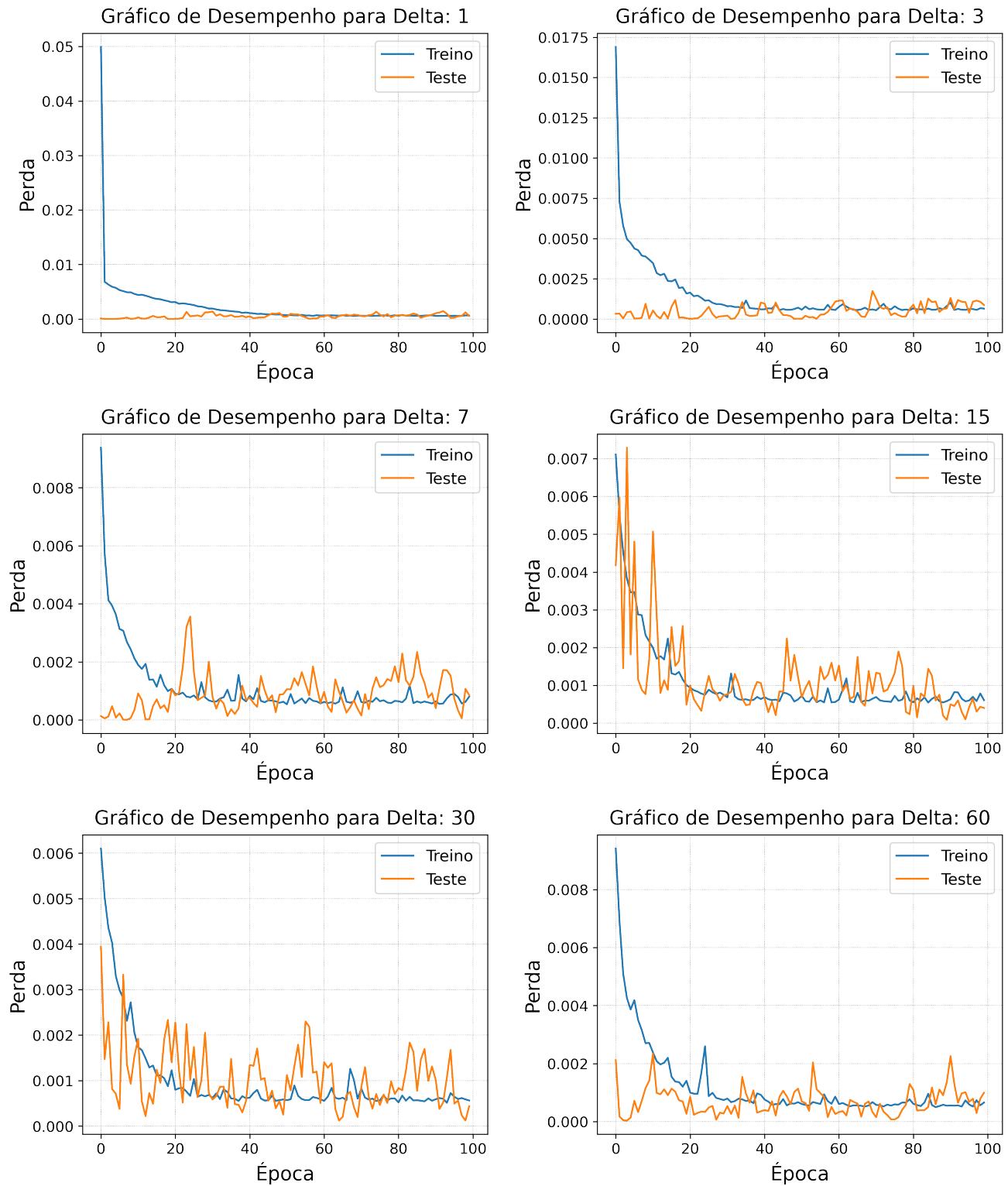
**A**PÓS a aplicação dos algoritmos realizou-se a análise dos valores obtidos. Para uma rede neural, as funções de perda são responsáveis por quantificar o desempenho em relação à aprendizagem com os dados de treinamento.

## Comparativo de Desempenho - Configuração 1



**Figura 12.** Gráfico com as curvas de aprendizado para os conjuntos de treinamento de teste, referentes à primeira configuração dos parâmetros da rede neural LSTM.

## Comparativo de Desempenho - Configuração 2



**Figura 13.** Gráfico com as curvas de aprendizado para os conjuntos de treinamento de teste, referentes à segunda configuração dos parâmetros da rede neural LSTM.

## Comparativo de Desempenho - Configuração 3



**Figura 14.** Gráfico com as curvas de aprendizado para os conjuntos de treinamento de teste, referentes à terceira configuração dos parâmetros da rede neural LSTM.

O erro é calculado com base nas previsões da rede, sendo que os erros de cada previsão são agregados e ponderados para obter a nota final da rede. Essa nota é utilizada para avaliar seu desempenho. Em tarefas de regressão, a função perda definida pelo erro quadrático médio, MSE, é frequentemente utilizada para obter valores reais como resultado da rede neural [14].

Pelas tabelas anteriores, pode-se perceber que os testes resultaram em um valor muito alto para o MSE. Geralmente, um valor muito elevado para essa métrica pode indicar uma estimativa de variação fortemente tendenciosa ou alta, ou mais provavelmente, uma combinação de ambos. Em alguns casos, isso pode sugerir a necessidade de uma modelagem mais refinada. Em outras situações, também é possível que não haja informações suficientes no conjunto de dados para que as inferências possam ser bem elaboradas.

Observando as Figuras 12, 13 e 14, pode-se visualizar os gráficos de desempenho do modelo de aprendizagem ao longo do tempo, para cada uma das configurações que foram utilizadas. É possível notar que o modelo apresenta anomalias relacionadas não somente ao *overfitting*. Para todos casos, observa-se que o conjunto de dados não apresenta representatividade, devido às lacunas e ruídos existentes entre as duas curvas<sup>31</sup>.

Um bom ajuste é o objetivo de qualquer algoritmo de aprendizado. Geralmente, ele é identificado por uma perda nos conjuntos de treinamento e de validação, a qual diminui para um ponto de estabilidade com um intervalo mínimo entre os dois valores de perda final. Embora os resultados obtidos para cada configuração dos hiperparâmetros utilizados pela rede neural estejam distorcidos devido aos valores existentes no conjunto de dados, pode-se perceber que, dentre as opções, aquela com o valor de  $\Delta t = 15$  é a que apresentou o melhor desempenho para o modelo de aprendizagem. Isso nos leva a discussão a respeito do teste de Friedman<sup>32</sup>.

Em tarefas de aprendizado de máquina, geralmente precisamos determinar se as amostras de dados possuem distribuições iguais ou diferentes. Uma das maneiras para se responder a esse questionamento é por meio da aplicação de testes de significância estatística, os quais podem quantificar a probabilidade das amostras terem a mesma distribuição. Caso os conjuntos de dados não tenham uma distribuição gaussiana familiar, deve-se recorrer às versões não paramétricas desses testes de significância, como por exemplo, o teste de Friedman.

Assim como os demais testes dessa categoria, o teste de Friedman retorna um valor  $p$  que pode ser utilizado para interpretar seu resultado. Este valor representa a probabilidade de se observar as amostras de dados, supondo-se que elas foram retiradas de uma população com a mesma distribuição. O valor  $p$  pode ser interpretado no contexto de um nível de significância  $\alpha = 0.05$ . Se o valor de  $p$  estiver abaixo desse nível de significância, o teste diz que há

<sup>31</sup>J. Brownlee. *How to use Learning Curves to Diagnose Machine Learning Model Performance*. Disponível em: <https://bit.ly/2MXSKUe>

<sup>32</sup>Wikipedia. *Friedman Test*. Disponível em: <https://bit.ly/3ft3XIE>

evidências suficientes para rejeitar a hipótese nula e que as amostras provavelmente foram extraídas de populações com distribuições diferentes, ou seja:

- $p \leq \alpha$ : rejeita  $H_0$ , as distribuições são diferentes;
- $p > \alpha$ : falha ao rejeitar  $H_0$ , as distribuições são iguais.

Para este artigo, aplicou-se o teste de Friedman sobre os valores das previsões obtidas, com o objetivo de verificar se eles são dependentes do passo de tempo escolhido (hipótese nula). A ideia é comparar os valores obtidos para cada  $\Delta t$ , de forma a comprovar se a sua escolha é importante para a predição do número de casos confirmados de COVID-19.

A aplicação do teste de Friedman para os valores das previsões realizadas para todos os  $\Delta t$  de cada configuração retornou o valor  $p = 0$ . Com isso, a hipótese nula foi rejeitada o que nos indica que pelo menos um dos valores utilizados como de passo de tempo apresentou um efeito diferente em nosso modelo de aprendizado.

Analizando as Figuras 12, 13 e 14 pode-se notar que o valor de  $\Delta t = 15$  é aquele que apresentou o melhor resultado de predição. Considerando que os órgãos responsáveis pelos setores de saúde informam que o período máximo para a manifestação dos sintomas de infecção pelo COVID-19 é de aproximadamente 14 dias<sup>33</sup>, é possível perceber que este passo de tempo tende a ser a melhor opção para a predição dos casos confirmados de infecção pelo coronavírus, independente da configuração de hiperparâmetros que foi utilizada, mesmo com a baixa representatividade do conjunto de dados.

## VIII. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho possibilitou entender como que uma técnica de aprendizado de máquina pode ser utilizada para realizar a análise preditiva de dados. Para isso, realizou-se um estudo sobre a evolução global do número de casos confirmados de COVID-19.

Para se atingir uma compreensão dessa realidade, foram coletados dados relacionados às ocorrências da doença em diversos países, juntamente com informações demográficas e socioeconômicas. O conjunto de dados foi preparado, de forma a eliminar possíveis inconsistências que pudessem prejudicar a análise por meio do algoritmo de aprendizado. Em seguida, aplicou-se uma técnica de aprendizado supervisionada, utilizando um modelo de regressão baseado em redes neurais artificiais do tipo LSTM.

Os conjuntos preditivos para o número de casos confirmados foram obtidos utilizando-se técnicas para a seleção dos atributos, as quais foram baseadas no método *Wrapper*, em conjunto com o Teorema de Takens. Cada um desses conjuntos foi avaliado por meio do cálculo de seu erro médio quadrático. A aplicação do teste de Friedman revelou que os valores obtidos para esses conjuntos são independentes do passo de tempo escolhido.

<sup>33</sup>World Health Organization. *Coronavirus Disease 2019 - Situation Report*. Disponível em: <https://bit.ly/2Bggp8W>

Analisando-se os comparativos de desempenho, pode-se perceber que o passo de tempo correspondente a um intervalo de 15 dias entre as amostras é o mais indicado para se prever um caso confirmado de COVID-19, independentemente dos ajustes que foram realizados nos hiperparâmetros da rede neural. Pelos gráficos de desempenho, percebeu-se que o modelo apresentou anomalias. Esse fato também foi evidenciado pelo alto valor obtido para o MSE.

Portanto, pode-se concluir que mesmo com a adoção de técnicas para a seleção dos atributos preditivos, o conjunto de dados não apresentou uma boa representatividade. Logo, em pesquisas futuras, é interessante que se utilize um conjunto de dados mais completo e com atributos mais significativos, que possam contribuir para gerar resultados mais coerentes. Com isso, seria possível verificar a real eficácia das técnicas utilizadas neste trabalho.

## Referências

- [1] I. Davidsohn. "Virus as Organism. Evolutionary and Ecological Aspects of Some Human Virus Diseases". *American Journal of Clinical Pathology*, vol. 16, pp. 662, Oct. 1946.
- [2] R. Shepost. "Coronavirus 2019 (COVID-19)". *Salem Press Encyclopedia of Health* [Online]. Disponível em: <https://bit.ly/2Ve2UnE>. Acessado em: 6 de abril de 2020.
- [3] S. J. Flint et al. "Principles of Virology". *Molecular Biology*, 3<sup>rd</sup> ed., vol. 1, Washington, WA, USA: ASM Press, 2008, 569 pp.
- [4] C. I. Paules; H. D. Marston; A. S. Fauci. "Coronavirus Infections - More Than Just the Common Cold". *JAMA*, vol. 323, no. 8, pp. 707-708, Jan. 2020, 10.1001/jama.2020.0757.
- [5] P. A. Morettin; C. M. C. Toloi. "Análise de Séries Temporais". 2<sup>nd</sup> ed., São Paulo, SP, BRA: Edgard Blücher Ltda., 2006, pp. 3.
- [6] R. L. P. Silva. "Aplicação de teoria de sistemas dinâmicos para inferência de causalidade entre séries temporais sintéticas e biológicas". Dissertação de Mestrado, Departamento de Física, Universidade Estadual Paulista, São Paulo, Brasil, 2018.
- [7] S. Hochreiter; J. Schmidhuber. "Long Short Term Memory". *Neural Computation*, vol. 9, pp. 1735-1780, Dec. 1997, 10.1162/neco.1997.9.8.1735.
- [8] A. De Myttenaere et al. "Mean absolute percentage error for regression models". *Neurocomputing*, vol. 192, pp. 38-48, Jun. 2016, 10.1016/j.neucom.2015.12.114.
- [9] P. N. Tan; M. Steinbach; V. Kumar. "Introdução ao Data Mining: Mineração de Dados". 1<sup>st</sup> ed., Rio de Janeiro, RJ, BRA: Editora Ciência Moderna Ltda., 2009, pp. 53-77.
- [10] X. Yan; X. Su. "Linear Regression Analysis: Theory and Computing". 1<sup>st</sup> ed., Singapore: World Scientific Publishing Co. Pte. Ltd., 2009, pp. 1-18.
- [11] F. Takens. "Detecting Strange Attractors in Turbulence". *Dynamical Systems and Turbulence*, vol. 898, pp. 366-381, Warwick, 1980.
- [12] L. Dos Santos; E. E. N. Macau. "Caracterização da Dinâmica Caótica em Séries Temporais". Workshop dos Cursos de Computação Aplicada do INPE, São José dos Campos, 2010.
- [13] G. L. Baker; J. P. Gollub. "Chaotic Dynamics: an Introduction". Cambridge University Press, 1998.
- [14] G. H. Lee. "Aprendizado de máquina profundo na análise de segmentação de clientes". Trabalho de Conclusão de Curso, Escola Politécnica da Universidade de São Paulo, São Paulo, Brasil, 2018.

• • •