



**UNIVERSIDADE DE BRASÍLIA
FACULDADE DE MEDICINA
PÓS-GRADUAÇÃO EM CIÊNCIAS MÉDICAS**

YURI ALMEIDA DE OLIVEIRA

**MORBILETALIDADE NA COVID-19: USO DE INTELIGÊNCIA ARTIFICIAL
PARA PREDIÇÃO DO DESFECHO CLÍNICO A PARTIR DE DADOS
CLÍNICOS DA ADMISSÃO HOSPITALAR**

BRASÍLIA

2024

YURI ALMEIDA DE OLIVEIRA

**MORBILETALIDADE NA COVID-19: USO DE INTELIGÊNCIA ARTIFICIAL
PARA PREDIÇÃO DO DESFECHO CLÍNICO A PARTIR DE DADOS
CLÍNICOS DA ADMISSÃO HOSPITALAR**

Dissertação de Mestrado
apresentada ao Programa de Pós-
graduação em Ciências Médicas da
Faculdade de Medicina,
Universidade de Brasília, como
requisito parcial à obtenção do título
de Mestre em Ciências Médicas.

Orientadora: Profa. Dra. Vânia Maria
Moraes Ferreira

BRASÍLIA

2024

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de ensino, estudo ou pesquisa, desde que citada a fonte.

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Am Almeida de Oliveira, Yuri
MORBILETALIDADE NA COVID-19: USO DE INTELIGÊNCIA
ARTIFICIAL PARA PREDIÇÃO DO DESFECHO CLÍNICO A PARTIR DE
DADOS CLÍNICOS DA ADMISSÃO HOSPITALAR / Yuri Almeida de
Oliveira; orientador Vânia Maria Moraes Ferreira;
co-orientador . -- Brasília, 2024.
77 p.

Tese(Mestrado em Ciências Médicas) -- Universidade de
Brasília, 2024.

1. Inteligência Artificial. 2. Aprendizado de Máquina. 3.
COVID-19. 4. Evolução Clínica. 5. Registros Eletrônicos de
Saúde. I. Moraes Ferreira, Vânia Maria, orient. II. , ,
co-orient. III. Título.

YURI ALMEIDA DE OLIVEIRA

**MORBILETALIDADE NA COVID-19: USO DE INTELIGÊNCIA ARTIFICIAL
PARA PREDIÇÃO DO DESFECHO CLÍNICO A PARTIR DE DADOS
CLÍNICOS DA ADMISSÃO HOSPITALAR**

Dissertação de Mestrado
apresentada ao Programa de Pós-
graduação em Ciências Médicas da
Faculdade de Medicina,
Universidade de Brasília, como
requisito parcial à obtenção do título
de Mestre em Ciências Médicas.

Orientadora: Profa. Dra. Vânia Maria
Moraes Ferreira

Data da defesa da dissertação: 29/05/2024

BANCA EXAMINADORA

Prof.^a Dr.^a Vânia Maria Moraes Ferreira
Universidade de Brasília

Dr. Túlio Frade Reis

Prof. Dr. Adson Ferreira da Rocha
Universidade de Brasília

Dr. Juliana Tessari Dias Rohr
Membro Suplente

Dedico à minha família, pelo amor e suporte.

Ao meu primo Fause, meu primeiro grande amigo.

À minha esposa, pelo companheirismo e incentivo.

AGRADECIMENTOS

À Universidade de Brasília, onde tudo começou.

Ao professor Hervaldo, por todo apoio desde o início da graduação e por ter aberto minha mente para a união entre a medicina e a tecnologia. Além disso, por ter coordenado, com maestria, o estudo de coorte que originou os dados utilizados neste trabalho.

À professora Vânia, por ter aceitado me orientar durante esta jornada. Mesmo fora de sua zona de conforto, agregou diversos conhecimentos e observações.

Ao professor Francisco de Assis, pelos seus ensinamentos e apoio nos testes dos classificadores e no pré-processamento dos dados.

Ao Programa de Pós-Graduação em Ciências Médicas (PPGCM) que, por meio de seus docentes, permitiu-me construir um arsenal de conhecimentos que muito contribuiu para a confecção desta obra científica.

Ao time de pesquisadores que participaram do estudo de coorte, coletando dados de extrema importância durante um dos momentos mais delicados da história da medicina.

A todos os demais, que fizeram parte desta jornada e ajudaram, direta ou indiretamente, na realização deste trabalho.

RESUMO

Introdução: Em 2019 a COVID-19 iniciou um dos maiores desafios sanitários da história, atingindo a condição de pandemia já no ano seguinte. De acordo com dados da Organização Mundial da Saúde (OMS), até outubro de 2023 havia 771.151.224 casos confirmados e 6.960.783 mortes diretas. O elevado número de casos simultâneos, aliado à grande quantidade de casos graves, levou ao colapso dos sistemas de saúde. Sistemas capazes de prever indivíduos com maior risco de evoluir para a forma grave da doença poderiam otimizar a alocação e direcionamento de recursos. Assim, diversos estudos foram publicados na tentativa de encontrar preditores baseados em inteligência artificial (IA) para, por meio de dados demográficos, clínicos, laboratoriais e de imagem médica, encontrar os indivíduos mais susceptíveis a evolução desfavorável pela doença. **Objetivos:** analisar o desempenho de algoritmos baseados em Aprendizado de Máquina (AM) como preditores de evolução durante internação hospitalar em pacientes com COVID-19, utilizando dados da admissão hospitalar. **Metodologia:** foram utilizados dados coletados em uma coorte prospectiva e multicêntrica de pacientes internados por COVID-19. Os dados da admissão hospitalar e do desfecho foram pré-processados, levando em consideração a disponibilidade dos dados entre os participantes do estudo e relevância clínica. Por meio de experimentação, diversos algoritmos baseados em IA foram aplicados e os melhores foram validados utilizando Validação Cruzada de Monte Carlo. **Resultados:** classificadores baseados em IA foram capazes de prever, com 80% de acurácia, quais indivíduos evoluiriam com desfecho desfavorável ou não. Destaca-se o *Random Forest Classifier*, com Área sob a Curva (AUC) de 91%, valores preditivos positivos e negativos de 81% e 79%, respectivamente, e sensibilidade e especificidade de 46% e 95%. Sem prejuízo significativo na acurácia, a Classificação por Vetor de Suporte (SVC) atingiu valor preditivo positivo de 87% e especificidade de 97%, adequado para situações em que uma pequena quantidade de falso positivos é preferível. **Conclusão:** é possível utilizar algoritmos baseados em AM para prever evolução clínica desfavorável durante a internação pela COVID-19 com acurácia satisfatória.

Palavras-Chave: inteligência artificial, aprendizado de máquina, COVID-19, classificação, evolução clínica, registros eletrônicos de saúde

ABSTRACT

Introduction: In 2019, COVID-19 began one of the greatest public health challenges in history, reaching pandemic status the following year. By October 2023, there were 771,151,224 confirmed cases and 6,960,783 direct deaths. The high number of simultaneous cases, coupled with a large number of severe cases, led to the collapse of health systems. Systems capable of predicting individuals at higher risk of progressing to severe forms of the disease could optimize the allocation and direction of resources. Thus, various studies have been published in an attempt to find predictors based on artificial intelligence that, through demographic, clinical, laboratory, and medical imaging data, could identify individuals most susceptible to unfavorable disease progression.

Objectives: to analyze the performance of Machine Learning (ML) based algorithms as predictors of clinical outcomes during hospitalization in COVID-19 patients, using hospital admission data. **Methodology:** data collected from a prospective, multicenter cohort of hospitalized COVID-19 patients were used. Admission and outcome data were pre-processed, taking into account the availability of data among study participants and clinical relevance. Through experimentation, various artificial intelligence based algorithms were applied, and the best were validated using Monte Carlo Cross-Validation. **Results:** artificial intelligence-based classifiers were successfully developed, achieving an 80% accuracy rate in predicting individuals' progression towards unfavorable outcomes. Random Forest Classifier stands out with an Area Under the ROC Curve (AUC) of 91%, and positive and negative predictive values of 81% and 79%, respectively. Additionally, it exhibited a sensitivity of 46% and a specificity of 95%. Without significant detriment to accuracy, the Support Vector Classification (SVC) achieved a positive predictive value of 87% and specificity of 97%, rates that are suitable for situations where a small number of false positives is preferable. **Conclusion:** it is possible to use machine learning-based algorithms to predict unfavorable clinical outcomes during hospitalization for COVID-19 with satisfactory accuracy.

Keywords: artificial intelligence, machine learning, COVID-19, classification, clinical outcome, electronic health record

LISTA DE FIGURAS

Figura 1. Exemplo de Máquina de Vetores de Suporte e o hiperplano de decisão	19
Figura 2. Outro exemplo de classificação utilizando MVS	20
Figura 3. Separação dos grupos utilizando árvore de decisão	21
Figura 4. Exemplo de árvore de decisão	22
Figura 5. Exemplo de <i>Random Forest</i> com 3 árvores de decisão	23
Figura 6. Exemplo de Análise Discriminante Linear	25
Figura 7. Projeção dos dados bidimensionais em um espaço unidimensional	26
Figura 8. Exemplo de rede neural com duas camadas ocultas	28
Figura 9. Funções de ativação, sigmoide e ReLU	30
Figura 10. Exemplo de rede neural densa, recebendo 28 variáveis na entrada e constituída de 8 unidades na primeira camada, 4 unidades na segunda e 1 unidade de saída.....	47
Figura 11. Número de pacientes para cada desfecho grave	49

LISTA DE TABELAS

Tabela 1. Amostras do exemplo fictício de casos de investigação de pneumonia bacteriana.....	27
Tabela 2. Dados coletados para cada participante do estudo.....	37
Tabela 3. Exemplo da operação lógica "OU" para agrupamento de variáveis .	40
Tabela 4. Categorização dos leucócitos, linfócitos, plaquetas e ureia	41
Tabela 5. Variáveis, imputação e quantidade de indivíduos em cada conjunto de dados	43
Tabela 6. Distribuição dos gêneros	50
Tabela 7. Distribuição das etnias	50
Tabela 8. Distribuição do estado civil	51
Tabela 9. Distribuição da renda familiar	51
Tabela 10. Distribuição do tipo de moradia	51
Tabela 11. Análise das variáveis demográficas numéricas: idade do paciente, número de cômodos na casa que habita e número de moradores que vivem na mesma casa	52
Tabela 12. Análise das variáveis numéricas com relação ao desfecho	53
Tabela 13. Variáveis categóricas.....	54
Tabela 14. Análise dos sinais de dispneia nas primeiras 24h de internação com relação ao desfecho.	57
Tabela 15. Análise do estado geral à admissão hospitalar	57
Tabela 16. Distribuição das variáveis utilizadas no classificador baseado em AM, após categorização.	59

Tabela 17. Distribuição das variáveis utilizadas no classificador baseado em AM após imputação	60
Tabela 18. Acurácia média por classificador (executado pela biblioteca <i>Lazy Predict</i>) e conjunto de dados.....	62
Tabela 19. Resultados de cada classificador após 5000 simulações, com Intervalo de Confiança de 95%	66

LISTA DE ABREVIATURAS E SIGLAS

COVID-19	Doença do Coronavírus de 2019
SARS-CoV-2	Coronavírus 2 da Síndrome Respiratória do Adulto
OMS	Organização Mundial da Saúde
IA	Inteligência Artificial
AM	Aprendizado de Máquina
ML	<i>Machine Learning</i>
MVS	Máquina de Vetores de Suporte
ADL	Análise Discriminante Linear
RNA	Redes Neurais Artificiais
AP	Aprendizado Profundo
ReLU	Unidade Linear Retificada
DG	Descida do Gradiente
PCR	Reação em Cadeia da Polimerase
UTI	Unidade de Terapia Intensiva
DPOC	Doença Pulmonar Obstrutiva Crônica
IVAS	Infecção de Vias Aéreas Superiores
NSH	<i>National Health Institute</i>
NCI	<i>National Cancer Institute</i>
AHA	<i>American Heart Association</i>
SpO₂	Saturação Periférica de Oxigênio
ROC	Característica de Operação do Receptor
AUC	Área Sob a Curva ROC
SVC	Classificador por Vetores de Suporte
NuSVC	SVC com controle do número de vetores de suporte

SUMÁRIO

1.	INTRODUÇÃO	15
1.1	PANDEMIA PELA COVID-19.....	15
1.2	USO DE INTELIGÊNCIA ARTIFICIAL NA COVID-19.....	16
1.3	ALGORITMOS DE IA	17
1.3.1	DEFINIÇÕES.....	17
1.3.2	AM CLÁSSICO	17
1.3.2.1	MÁQUINA DE VETORES DE SUPORTE.....	18
1.3.2.2	ÁRVORES DE DECISÃO E RANDOM FOREST	20
1.3.2.3	ANÁLISE DISCRIMINANTE LINEAR	24
1.3.3	REDES NEURAIS ARTIFICIAIS E APRENDIZADO PROFUNDO .	26
2.	OBJETIVOS	34
2.1	OBJETIVO GERAL.....	34
2.2	OBJETIVOS ESPECÍFICOS	34
3.	MATERIAIS E MÉTODOS.....	35
3.1	COLETA DE DADOS	35
3.1.1	DESENHO DO ESTUDO	35
3.1.2	POPULAÇÃO	35
3.1.3	CRITÉRIOS DE INCLUSÃO E EXCLUSÃO	36
3.1.4	DADOS CATALOGADOS	36
3.2	PRÉ-PROCESSAMENTO DOS DADOS.....	39

3.3	CONSTRUÇÃO DOS CONJUNTOS DE DADOS	42
3.4	DESENVOLVIMENTO DO CLASSIFICADOR.....	44
4.	RESULTADOS	49
4.1	ANÁLISE DE DADOS DA COORTE: DESFECHOS	49
4.2	ANÁLISE DE DADOS DA COORTE: ANÁLISE DEMOGRÁFICA ..	50
4.3	ANÁLISE GLOBAL DAS VARIÁVEIS UTILIZADAS NO ESTUDO .	52
4.4	RESULTADOS DOS CLASSIFICADORES	61
5.	DISCUSSÃO	67
6.	CONCLUSÕES	73
7.	REFERÊNCIAS BIBLIOGRÁFICAS.....	74

1. INTRODUÇÃO

1.1 PANDEMIA PELA COVID-19

Iniciada em 2019, a Doença do Coronavírus (ou COVID-19), infecção respiratória causada pelo Coronavírus 2 da Síndrome Respiratória do Adulto (SARS-CoV-2), espalhou-se rapidamente pelo mundo, atingindo a condição de pandemia já em março de 2020 (1,2). De acordo com a Organização Mundial da Saúde (OMS), até 4 de outubro de 2023, já havia 771.151.224 casos confirmados e 6.960.783 mortes em todo o globo (3). A COVID-19 se caracteriza, principalmente, por um quadro de pneumonia, com febre, fadiga e tosse seca. No entanto, outros sintomas podem ser observados, como alterações gastrointestinais, anosmia ou alterações oftalmológicas (4). Quadros graves requerem cuidado intensivo, com necessidade de intubação orotraqueal e ventilação mecânica, levando à morte em casos extremos (4–7).

O aumento acelerado do número de infecções evidenciou o despreparo dos sistemas de saúde, afetando até mesmo países desenvolvidos. Houve falta de profissionais qualificados (8,9), de medicações (10), equipamentos (11) e reagentes para exames laboratoriais (12,13). Para lidar com as deficiências de recursos, observou-se o surgimento de novas estratégias, com mudanças nas políticas de gestão (1,2,8,9), distribuição de vacinas (14) o desenvolvimento de técnicas alternativas para tratamentos e diagnósticos (10,12,13,15). Os impactos sociais também foram notáveis, como o agravamento da situação econômica, aumento nos casos de violência doméstica e descontinuidade no tratamento de outras doenças (1).

1.2 USO DE INTELIGÊNCIA ARTIFICIAL NA COVID-19

Esta crise global e suas consequências devastadoras, bem como a escassez de recursos, motivou o desenvolvimento de diversas ferramentas baseadas em inteligência artificial (IA). Foram implementados algoritmos para auxiliar no diagnóstico da COVID-19, interpretação de exames de imagem, previsão da variação do número de casos, desenvolvimento de medicações e vacinas e previsão do prognóstico da doença em pacientes infectados (12,13,16–23).

O desenvolvimento de bons preditores de morbiletalidade durante uma pandemia poderia auxiliar no planejamento estratégico dos sistemas de saúde, indicando com antecedência, por exemplo, a necessidade de internação de um grupo de pacientes infectados (19). A qualidade dos dados utilizados para construir um preditor é fundamental, já que dados clínicos ruidosos, incompletos ou mal processados podem introduzir vieses que inviabilizam sua implementação (18).

Sendo a primeira grande pandemia da era digital, uma quantidade enorme de informações foi coletada (18,21). Com o objetivo de identificar quais os fatores de risco estão mais associados à gravidade da COVID-19, dezenas de estudos de coorte, estudos de caso-controle e séries de casos foram conduzidos (5–7). A partir destas informações, preditores de mortalidade e morbidade, baseados em IA, podem ser desenvolvidos (19).

Na tentativa de elaborar um preditor confiável e passível de ser utilizado na gestão pública em saúde, foi conduzido um estudo de coorte prospectivo e multicêntrico, para acompanhar pacientes internados por suspeita de COVID-19, a partir de dados clínicos e demográficos coletados entre dezembro de 2019 e janeiro de 2021. Por meio destes dados foi possível implementar algoritmos baseados em AM supervisionada e construir classificadores com desempenho satisfatório.

1.3 ALGORITMOS DE IA

1.3.1 DEFINIÇÕES

IA é um termo que vem sendo utilizado amplamente na literatura, mas frequentemente com definições distintas. Isto se torna mais evidente quando este termo é utilizado em conjunto com Aprendizado de Máquina (AM) ou *Machine Learning* (ML). Apesar de não haver um consenso, é seguro definir IA como um sistema que busca imitar a inteligência humana. Já AM é considerado um subconjunto da IA e diz respeito aos sistemas capazes de aprender por meio de conjuntos de dados, sem uma programação explícita sobre a forma com que deve tomar as decisões (24). Outro conceito frequentemente utilizado é o de Redes Neurais Artificiais (RNA), sistemas inspirados pelo funcionamento do sistema nervoso e capazes de aprender com os dados apresentados. As RNA podem apresentar diversas arquiteturas diferentes e podem ser definidas como um subconjunto do AM (20,25).

Existem diversos métodos de AM, cada um mais adequado para um tipo de tarefa. O aprendizado pode ser supervisionado, semi-supervisionado ou não supervisionado e as tarefas incluem classificação, agrupamento, regressão, localização, entre outros. Em uma tarefa de classificação, como no caso de previsão de morbiletalidade pela COVID-19 a partir de dados clínicos, diversos algoritmos supervisionados podem ser utilizados. Estes algoritmos recebem o nome de supervisionados pois os dados clínicos são apresentados rotulados, ou seja, em conjunto com o desfecho de cada paciente (26).

1.3.2 AM CLÁSSICO

Certos métodos de AM são frequentemente denominados clássicos para diferenciá-los das redes neurais artificiais profundas (ou Aprendizado Profundo, *Deep Learning*). Isto porque, habitualmente, se tratam de algoritmos mais leves, exigem menos esforço computacional para execução e treinamento e, assim, puderam ser largamente aplicados nas décadas passadas. Se enquadram nesta

categoria a regressão logística, árvores de decisão, máquina de vetor de suporte, vizinhos mais próximos (ou *Nearest Neighbors*), entre outros (27).

Cada método possui uma abordagem diferente para a análise de dados, mas todos trabalham encontrando parâmetros que minimizam o erro de classificação. Estes parâmetros são o aprendizado adquirido pela análise dos dados (26). A título de exemplificação de como máquinas são capazes de aprender através dos dados, alguns algoritmos serão descritos em maiores detalhes.

1.3.2.1 MÁQUINA DE VETORES DE SUPORTE

A Máquina de Vetores de Suporte (MVS) é um algoritmo de AM desenvolvido por Vladimir Vapnik e colaboradores. Ela funciona encontrando um hiperplano capaz de dividir um conjunto de dados em grupos ou classes (28,29). Para entender este processo pode-se pensar em um conjunto de dados clínicos como uma nuvem de pontos em um espaço n-dimensional. A título de exemplo, considere um problema fictício de diagnóstico de pneumonia bacteriana com base apenas na temperatura corporal da admissão e número de leucócitos. Neste problema, os dados de entrada se apresentam em apenas duas dimensões (temperatura e leucócitos) e duas saídas são possíveis, presença de pneumonia bacteriana ou não. É possível registrar os dados clínicos em um gráfico, onde o eixo x representa o número de leucócitos/mm³ e o eixo y a temperatura corporal. Neste gráfico, pode-se representar os pontos com cores diferentes, em vermelho os pontos onde a pneumonia bacteriana foi confirmada pelo exame padrão ouro, em azul aqueles pacientes que não tinham a doença e em verde dados de pacientes novos, cujo diagnóstico não foi confirmado. A **Figura 1** representa este exemplo.

Na **Figura 1** é possível se observar o hiperplano de decisão (linha tracejada), que separa, de forma linear, as duas classes. É evidente a presença de classificações incorretas, como um ponto vermelho à esquerda do hiperplano representando um caso de pneumonia sem febre e sem leucocitose, mas este hiperplano foi definido pela função linear que minimiza os erros. Neste exemplo,

caso se queira diagnosticar os pontos verdes somente pela temperatura e número de leucócitos, basta determinar se estão abaixo (ou à esquerda) ou acima (ou à direita) do hiperplano, o diagnóstico será negativo ou positivo para pneumonia, respectivamente.

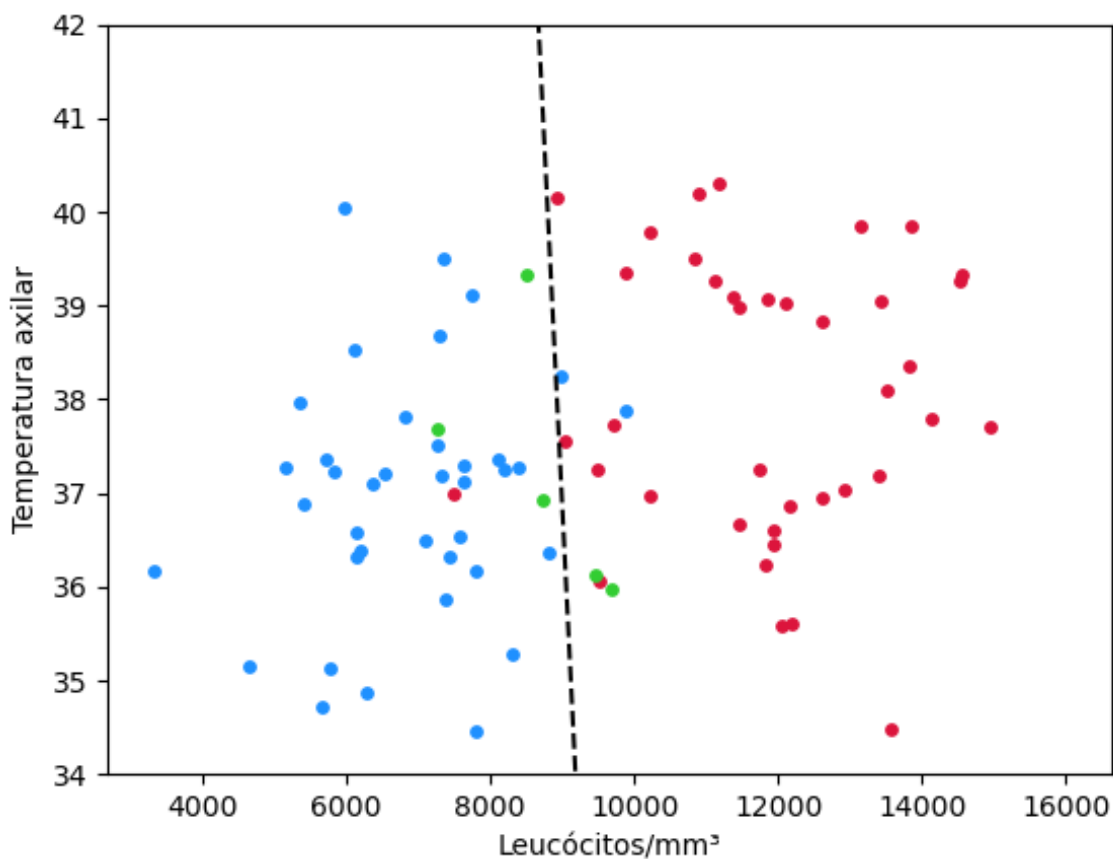


Figura 1. Exemplo de Máquina de Vetores de Suporte e o hiperplano de decisão.
Fonte: Autor

Em um mesmo conjunto de dados, diversos hiperplanos podem ser usados para separar as diferentes classes. No exemplo anterior é possível imaginar diversas linhas que poderiam separar os dois grupos. Para definir, entre os diferentes hiperplanos possíveis, qual o mais adequado, uma MVS busca encontrar aquele que maximiza a margem entre os grupos. Esta margem é definida utilizando vetores de suporte, que dão nome ao algoritmo. Os vetores de suporte são os pontos mais próximos ao hiperplano de decisão e, portanto, os mais difíceis de classificar. A partir desses pontos, as margens são definidas e o hiperplano que maximiza a distância entre as margens é traçado. A **Figura 2**

traz um novo conjunto de dados, mas seguindo o mesmo exemplo de diagnóstico de pneumonia, para representar os vetores de suporte (pontos circulados), margens (linhas pontilhadas) e hiperplano de decisão (linha contínua).

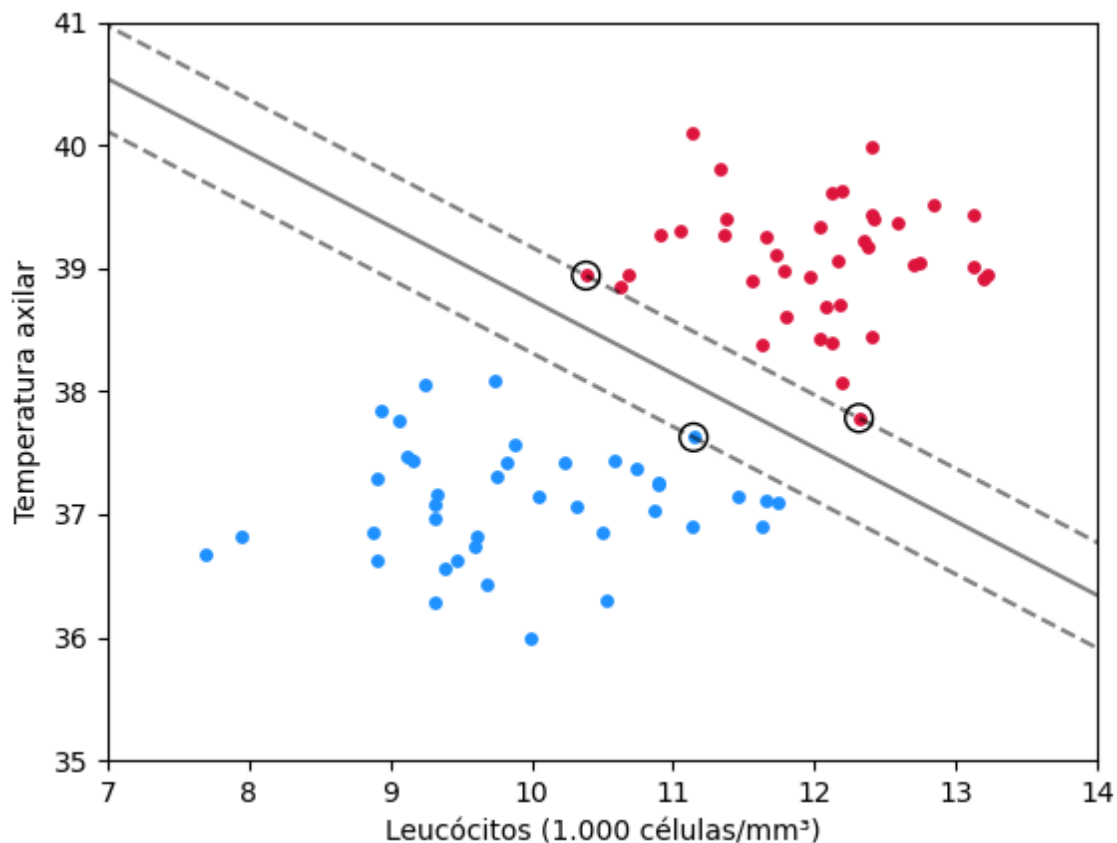


Figura 2. Outro exemplo de classificação utilizando MVS. Desta vez um conjunto de dados com maior separação foi escolhido para representar as margens (linha tracejada) e vetores de suporte (pontos circulados). Fonte: Autor

1.3.2.2 ÁRVORES DE DECISÃO E *RANDOM FOREST*

Árvores de decisão são algoritmos que se baseiam na definição de regras para classificar um conjunto de dados (30). Para entender melhor o seu funcionamento, pode-se utilizar o mesmo exemplo descrito na sessão 1.3.2.1, mais especificamente os dados plotados na **Figura 2** para diagnóstico de pneumonia bacteriana. Neste exemplo pode-se afirmar que, caso a temperatura axilar esteja acima de 38,2°C, com certeza se trata de um caso de pneumonia bacteriana. Pode-se então definir esta faixa de temperatura como a primeira regra para separação do conjunto de dados, dividindo-os em um grupo apenas

de pacientes com pneumonia bacteriana (denominada de nó A) e outro que contem pacientes com e sem a doença (denominada de nó B). O nó B poderá ser dividido novamente, desta vez utilizando a contagem de leucócitos como regra. Caso a contagem esteja acima de $12.000/\text{mm}^3$ pode-se classificar como pneumonia bacteriana, do contrário, outro diagnóstico. Assim, o nó B será dividido em dois novos nós, C e D, que separam corretamente todos os pacientes. A **Figura 3** exemplifica essa separação, com a linha azul representando a separação pela temperatura axilar (com ponto de corte em $38,2^\circ\text{C}$) e a linha verde representando a separação com base na contagem de leucócitos (com ponto de corte em $12.000/\text{mm}^3$). Pode-se observar que as classes são perfeitamente separadas desta forma.

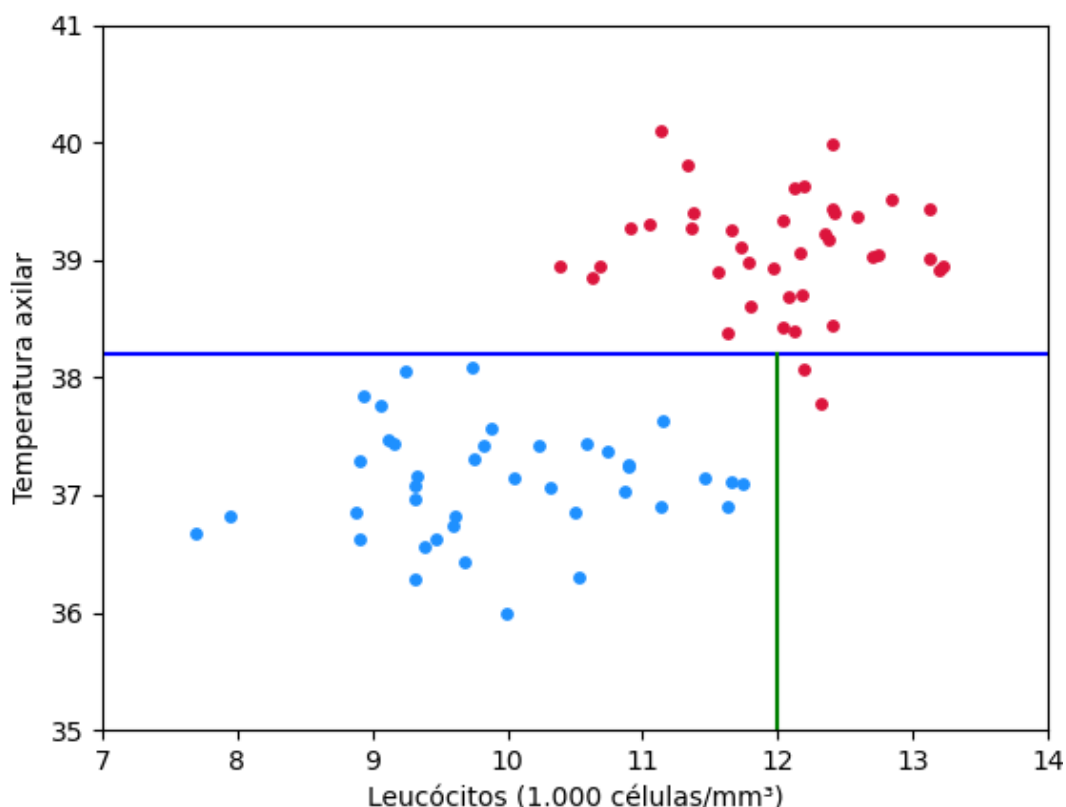


Figura 3. Separação dos grupos utilizando árvore de decisão. A linha azul representa a separação dos dados com base na temperatura axilar ($38,2^\circ\text{C}$) e a linha verde representa a separação com base no número de leucócitos ($12.000/\text{mm}^3$). Fonte: Autor

Neste exemplo, os pontos de corte foram definidos manualmente, mas os pontos de corte podem ser aprendidos com algoritmos de AM (31). Este aprendizado é fundamental para a aplicabilidade do algoritmo, já que casos reais tendem a ser muito complexos e utilizam dezenas de variáveis.

O nome árvore de decisão vem do fato de que a cadeia de regras utilizadas para separar os dados pode ser imaginada como a estrutura de um galho, enquanto a classificação atribuída a cada dado pode ser vista como uma folha. A **Figura 4** exemplifica esta estrutura.

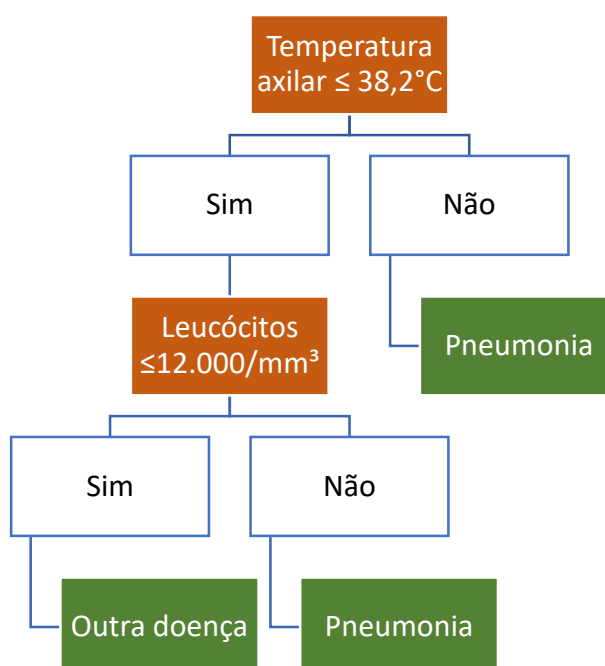


Figura 4. Exemplo de árvore de decisão. Os retângulos laranja representam as regras, enquanto os verdes representam as classes. Fonte: Autor

Já o *Random Forest* é um algoritmo classificado como *ensemble*. Isso porque ele utiliza diferentes árvores de decisão para chegar a um resultado. Este algoritmo separa o conjunto de dados de treinamento, de forma aleatória, em diferentes subconjuntos. Cada subconjunto será utilizado para treinar uma nova árvore de decisão. Uma vez definidas as diferentes árvores, a classificação será realizada por meio de uma votação. Cada árvore irá receber os mesmos dados de entrada (no nosso exemplo, os dados de um mesmo paciente) e votará a qual classe (presença de pneumonia bacteriana ou não) aqueles dados pertencem. A

classe com maior quantidade de votos será a definida (30). A **Figura 5** exemplifica o resultado de um *Random Forest* com 3 árvores de decisão diferentes, utilizando o mesmo exemplo anterior.

Considere um paciente com temperatura axilar de 38°C e contagem de leucócitos de 12.000/mm³. Utilizando o exemplo de *Random Forest* da **Figura 5**, tem-se, para a árvore à esquerda, o diagnóstico de pneumonia. Na árvore central o diagnóstico é de “outra doença” e pela árvore à direita o diagnóstico também seria de pneumonia bacteriana. Como pneumonia recebeu 2 dos 3 votos, este seria o diagnóstico final deste paciente.

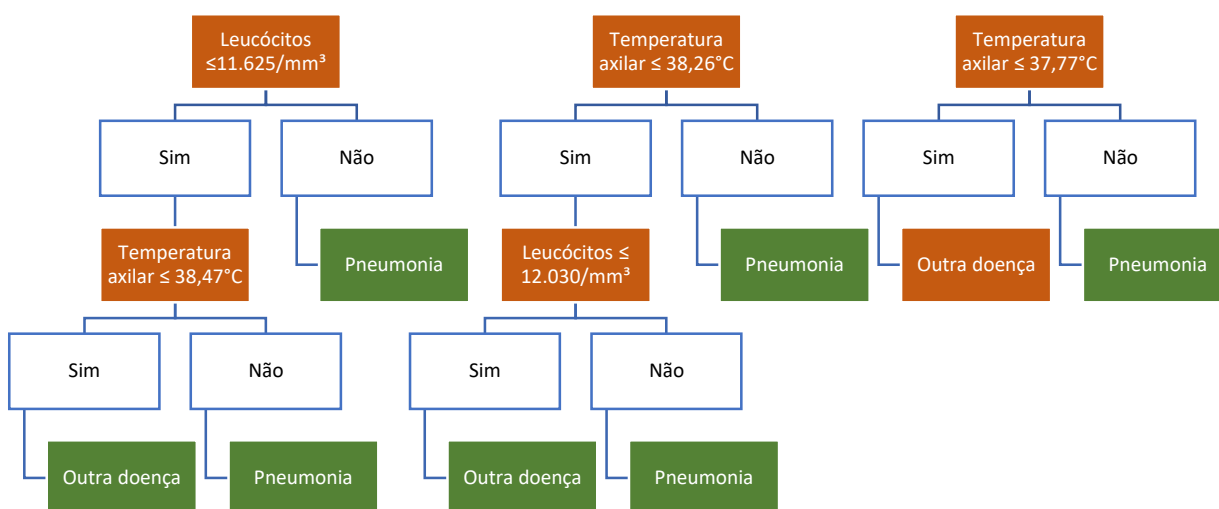


Figura 5. Exemplo de *Random Forest* com 3 árvores de decisão. Fonte: Autor

O *Random Forest* permite um bom desempenho na classificação por ser menos propenso ao sobreajuste (ou *overfitting*) e, assim, melhor desempenho com novos dados (que não foram usados para o treinamento) (30).

1.3.2.3 ANÁLISE DISCRIMINANTE LINEAR

A Análise Discriminante Linear (ADL) funciona reduzindo a dimensionalidade dos dados, de modo a aumentar a separabilidade entre as classes (32). Reduzir a dimensionalidade significa representar cada ponto, em nosso exemplo cada paciente, utilizando uma quantidade menor de variáveis (apenas um número ao invés da temperatura e contagem de leucócitos). Isto acontece por meio de 3 passos. O primeiro consiste em definir a separabilidade entre as classes calculando a diferença entre a média. Ainda utilizando o exemplo do diagnóstico de pneumonia bacteriana da sessão 1.3.2.1, com os mesmos dados da **Figura 2**, tem-se uma média de 9.910 leucócitos/mm³ no grupo sem pneumonia bacteriana, 12.008 leucócitos/mm³ no grupo com a doença e 10.959 leucócitos/mm³ considerando todos os pacientes. A temperatura axilar média é de 37,09°C no grupo sem pneumonia, 39,06°C no grupo com a doença e 38,08°C considerando todos os pacientes. A diferença entre as médias de cada classe para a média geral (de todos os pacientes) é chamada variância entre classes.

O segundo passo consiste em calcular a variância interclasse, ou seja, a distância de cada ponto para a média da classe a qual pertence. Já o terceiro passo é de fato a redução da dimensionalidade dos dados, maximizando a variância entre classes e reduzindo a variância interclasses. Este algoritmo assume que os dados são linearmente separáveis, ou seja, é possível separar os pontos através de uma função linear, que os dados têm distribuição Gaussiana e que as matrizes de covariância das diferentes classes são iguais. A partir destes pressupostos, a ADL é capaz de implementar um conjunto de equações e determinar a redução da dimensionalidade (32).

Este processo complexo pode ser exemplificado por meio da **Figura 6** a seguir. Nela pode-se observar a mesma distribuição dos exemplos anteriores, com os pontos em duas dimensões (contagem de leucócitos no eixo x e temperatura axilar no eixo y). No entanto, foi determinada uma reta (em azul) que cruza o gráfico e estes pontos são então projetados sobre esta reta. As projeções dos pontos nesta reta agora podem ser vistas como dados unidimensionais, conforme demonstrado pela **Figura 7**. Nesta figura pode-se

observar as projeções dos dados em um espaço unidimensional, onde cada ponto possui apenas um valor x . Este valor x representa a combinação da contagem de leucócitos e temperatura axilar e, apesar de não ter valor clínico, pode ser usado, neste exemplo, para o diagnóstico da pneumonia. Pontos com valores superiores a 26,1 podem ser classificados como pertencentes a pacientes com pneumonia bacteriana, enquanto os pontos com valor inferior são de pessoas sem esta doença.

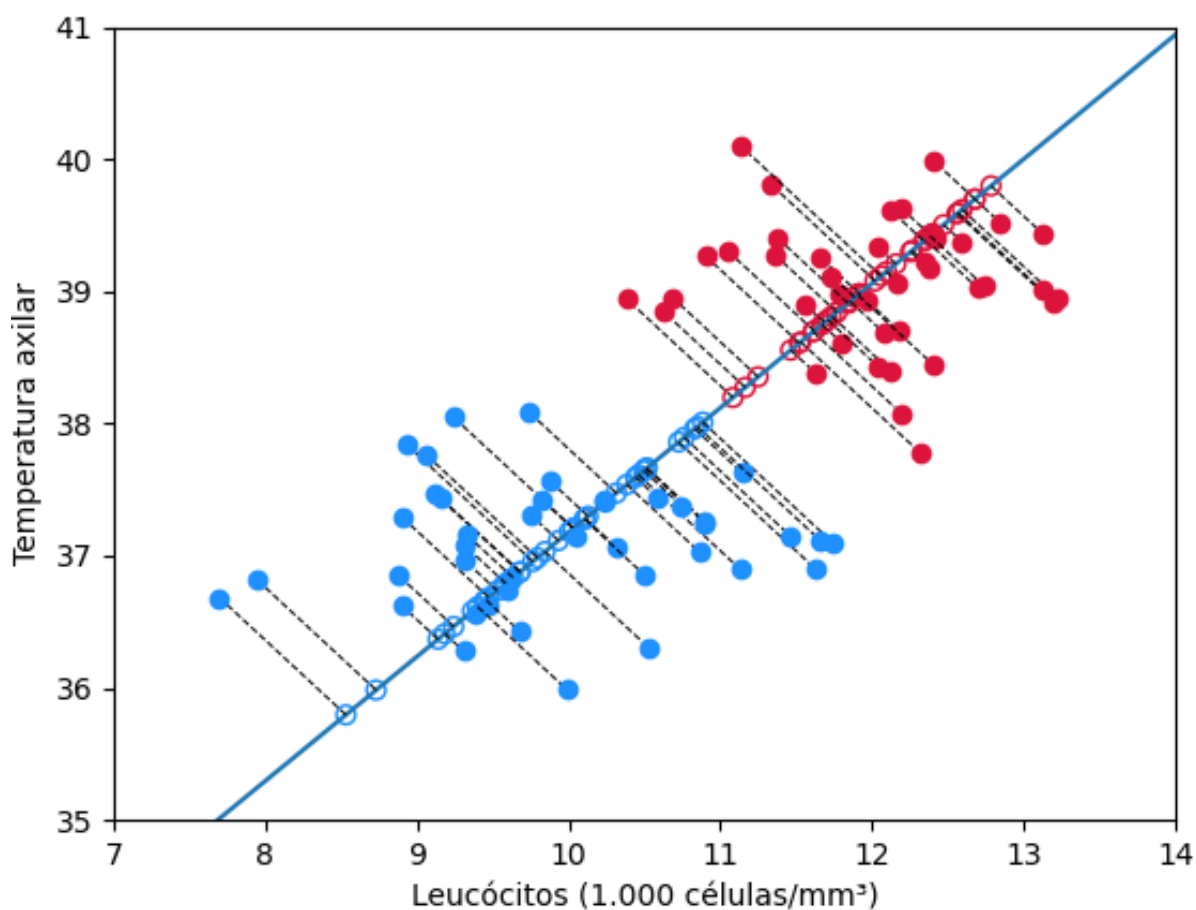


Figura 6. Exemplo de Análise Discriminante Linear. Neste exemplo os dados bidimensionais são projetados em uma reta, reduzindo-os para uma dimensão. Fonte: Autor

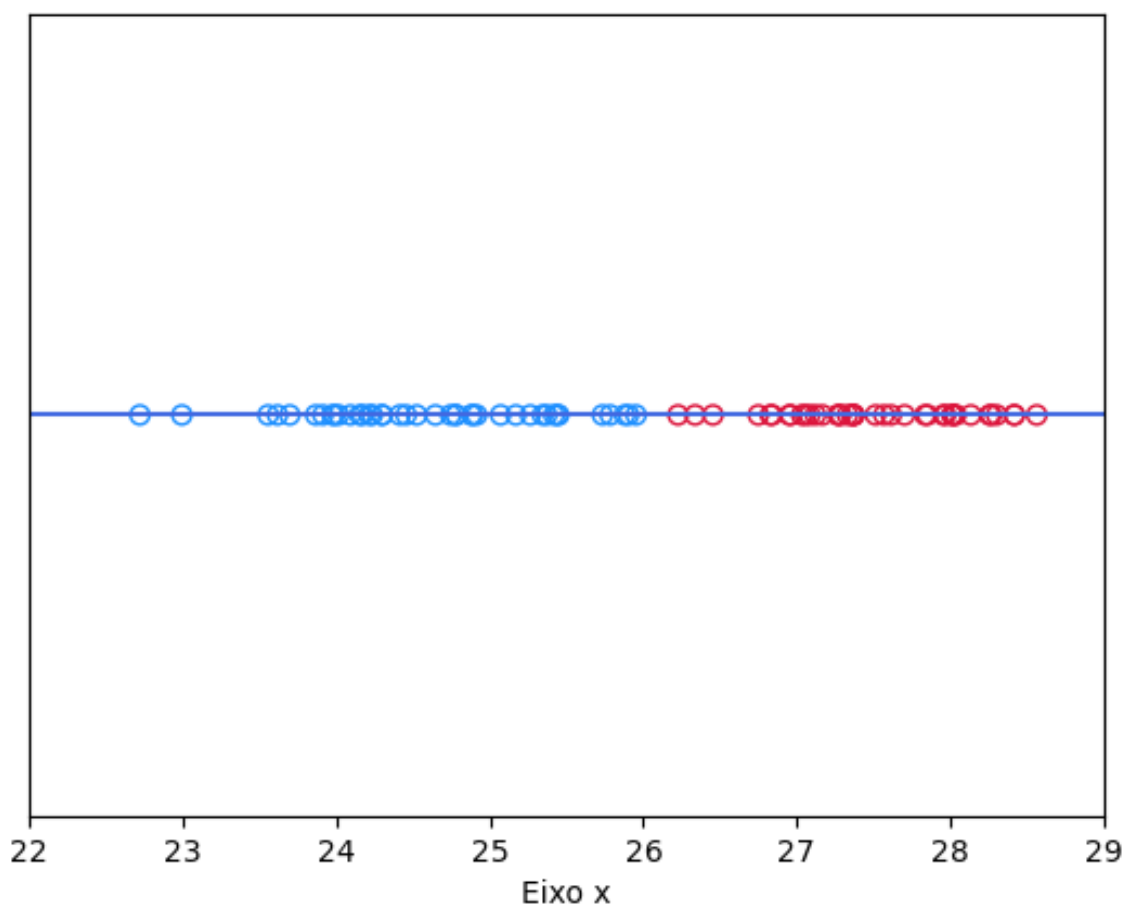


Figura 7. Projeção dos dados bidimensionais em um espaço unidimensional. Pode-se, por exemplo, classificar como portadores de pneumonia bacteriana apenas aqueles com valor acima de 26,10 (pontos vermelhos). Fonte: Autor

1.3.3 REDES NEURAIS ARTIFICIAIS E APRENDIZADO PROFUNDO

Atualmente as redes neurais artificiais têm aumentado sua popularidade, pelo aumento de poder de processamento dos computadores e aumento das informações disponíveis em forma digital. São algoritmos inspirados no funcionamento do sistema nervoso, onde uma rede de neurônios artificiais é capaz de receber os dados de entrada, como resultados de exames laboratoriais, e entregar, na saída, a qual classe o indivíduo que coletou os exames pertence, como a presença ou ausência de doença. Este método faz uso de sucessivas transformações não lineares para determinar se cada unidade (ou neurônio artificial) da rede será ativado ou não, simulando uma despolarização neuronal (33–36).

Para exemplificar o funcionamento de uma rede neural, serão utilizados dados de exemplos anteriores, sobre diagnóstico de pneumonia bacteriana por meio da temperatura axilar e contagem de leucócitos. Uma amostra de 10 pacientes está representada na **Tabela 1** a seguir. Neste exemplo os exames foram realizados durante a investigação clínica e o diagnóstico final foi determinado utilizando o exame padrão ouro.

Tabela 1. Amostras do exemplo fictício de casos de investigação de pneumonia bacteriana. É atribuído o valor 1 para os casos que foram confirmados como sendo pneumonia bacteriana e 0 para os casos em que a hipótese da doença foi descartada.

Paciente	Leucócitos (1.000/mm ³)	Temperatura axilar (°C)	Pneumonia bacteriana
1	12,41	40,0	Sim (1)
2	12,69	39,0	Sim (1)
3	8,88	36,8	Não (0)
4	11,96	38,9	Sim (1)
5	11,79	38,6	Sim (1)
6	9,73	38,0	Não (0)
7	10,74	37,4	Não (0)
8	12,11	39,6	Sim (1)
9	10,68	38,9	Sim (1)
10	10,90	37,2	Não (0)

Estes dados podem ser, então, utilizados para treinar uma rede neural artificial cuja topologia está representada na **Figura 8**. Nesta topologia observa-se uma camada de entrada (os dados clínicos da temperatura e contagem de leucócitos), uma camada “A” contendo 3 unidades (ou neurônios), outra camada “B” contendo 2 unidades e uma última camada “S” contendo uma unidade. As camadas com 3 e 2 unidades são chamadas de camadas ocultas, enquanto a última camada é chamada camada de saída. O nome camada oculta advém do fato de que os valores que estas camadas computam estão “escondidos” do usuário, que observa apenas os dados que entram no sistema (dados clínicos) e os de saída (diagnóstico).

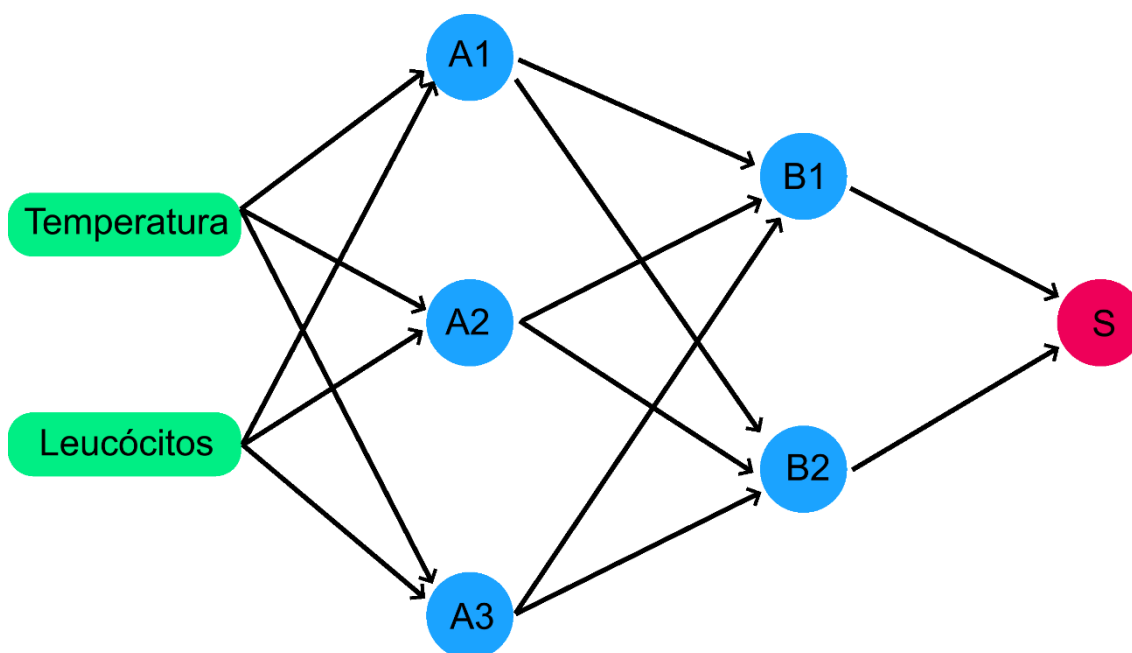


Figura 8. Exemplo de rede neural com duas camadas ocultas (camadas A e B).
Fonte: Autor

A **Figura 8** demonstra a conexão entre as diferentes unidades, representadas pelas setas. As unidades de uma camada n se conectam com todas as unidades da camada seguinte, $n + 1$. Pode-se pensar nestas conexões como axônios, carregando uma informação para a unidade seguinte. A informação que é passada de uma unidade para a próxima é o valor de saída desta unidade multiplicado por um coeficiente (ou peso) w , que é específico para

cada conexão. Por exemplo, a unidade A1 (primeira unidade da camada que terá a denominação de “A”) receberá dois valores diferentes, que são somados. Ao final, será adicionado um terceiro valor b , chamado viés ou bias, que é específico da unidade recebendo o sinal. Assim, o valor final z_1 que a unidade A1 receberá será o descrito na **Equação 1** abaixo.

Equação 1. Cálculo do valor z_1 recebido pela unidade A1

$$z_1 = \textit{Temperatura} * w_1 + \textit{Leucócitos} * w_2 + b_1$$

O mesmo acontecerá com as unidades A2 e A3, recebendo, respectivamente, os valores z_2 e z_3 . Cabe ressaltar que, apesar de também utilizarem os mesmos valores de temperatura axilar e contagem de leucócitos, z_1 , z_2 e z_3 serão valores distintos, pois serão calculados utilizando pesos (w) e vieses (b) diferentes.

Antes de cada unidade da camada A emitir seu respectivo valor z à camada B, este valor será transformado, passando por uma função não linear, chamada de função de ativação ($g(z)$). Esta etapa é extremamente importante, por ser a garantia de que a adição de novas camadas contribui para o aprendizado. Isto porque aplicar a operação linear $z_1 = w_1 * x + b_1$, sendo x uma variável hipotética qualquer, e aplicar uma nova transformação onde $z_2 = w_2 * z_1 + b_2$ é o mesmo que encontrar os valores w_3 e b_3 que calculam z_2 diretamente em função de x ($z_2 = w_3 * x + b_3$). Isto se deve à duas propriedades que definem transformações lineares, aditividade e homogeneidade.

Funções de ativação são então transformações não lineares que recebem o valor z calculado em uma unidade e retorna o valor a , que representa o quanto aquela unidade foi ativada pelos valores de entrada. Dois exemplos destas funções são a função sigmoide, definida pela **Equação 2**, e a Unidade Linear Retificada (ReLU), definida pela **Equação 3** (36). Os gráficos destas funções estão representados pela **Figura 9**.

Equação 2. Função sigmoide

$$a = \frac{1}{1 + e^{-z}}$$

Equação 3. Função ReLU

$$a = 0, \text{ se } z \leq 0;$$

$$a = z, \text{ se } z > 0.$$

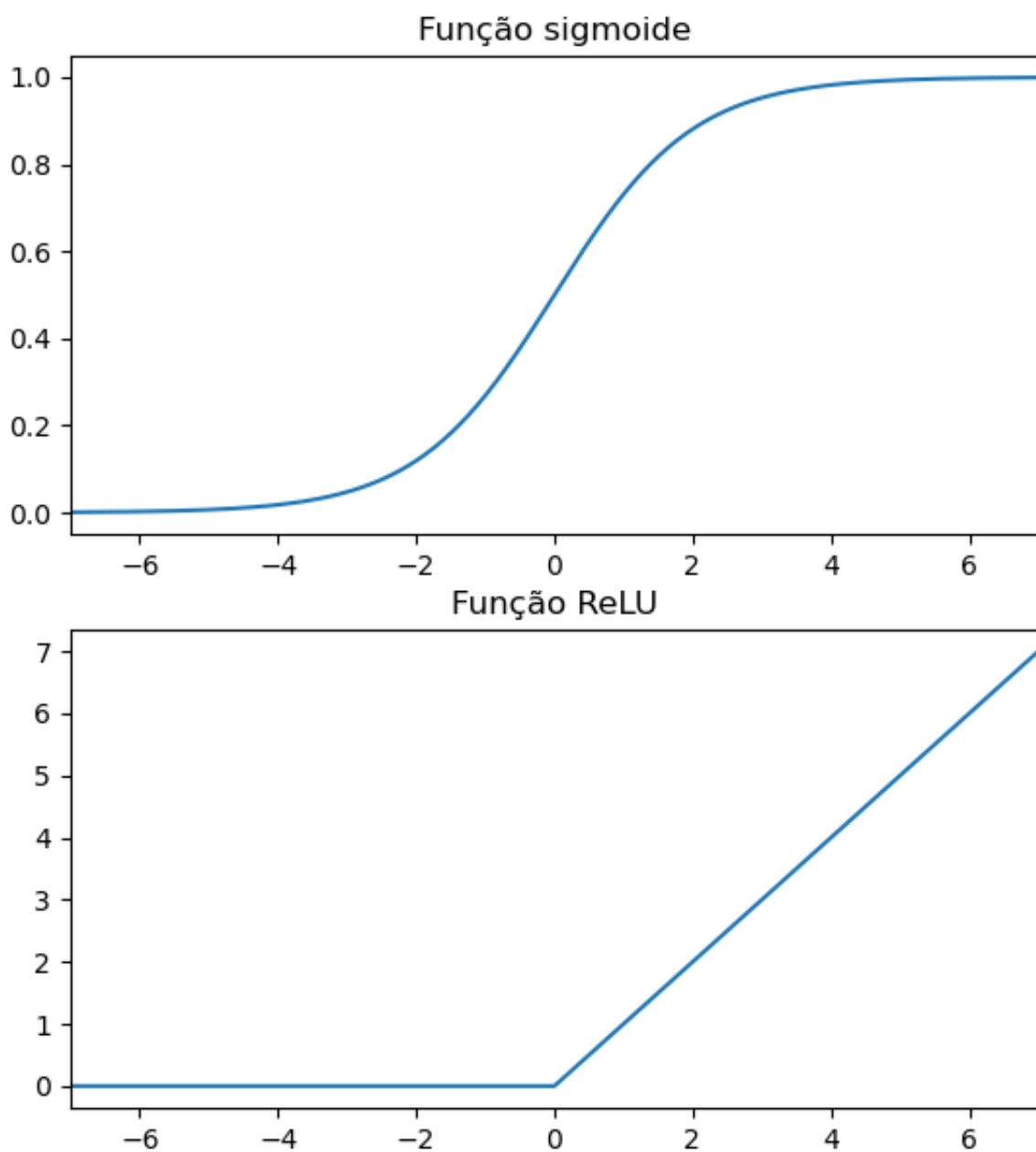


Figura 9. Funções de ativação, sigmoide e ReLU. Fonte: Autor

Será então o valor a que será transmitido à próxima camada, de modo que a unidade B1, por exemplo, receberá os valores a_1 , a_2 e a_3 , oriundos, respectivamente, das unidades A1, A2 e A3. Estes valores a serão então transmitidos à camada seguinte, seguindo os mesmos passos descritos previamente. A **Equação 4** descreve toda a computação que é realizada.

Equação 4. Cálculos realizados durante a propagação direta

Camada A:

$$z_1 = Temperatura * w_1 + Leucócitos * w_2 + b_1$$

$$z_2 = Temperatura * w_3 + Leucócitos * w_4 + b_2$$

$$z_3 = Temperatura * w_5 + Leucócitos * w_6 + b_3$$

$$a_1 = g(z_1)$$

$$a_2 = g(z_2)$$

$$a_3 = g(z_3)$$

Camada B:

$$z_4 = a_1 * w_7 + a_2 * w_8 + a_3 * w_9 + b_4$$

$$z_5 = a_1 * w_{10} + a_2 * w_{11} + a_3 * w_{12} + b_5$$

$$a_4 = g(z_4)$$

$$a_5 = g(z_5)$$

Camada de saída:

$$z_6 = a_4 * w_{13} + a_5 * w_{14} + b_6$$

$$a_6 = g(z_6)$$

Este caminho que parte dos valores de temperatura e contagem de leucócitos e chega ao valor a_6 , descrito na **Equação 4**, é chamado de Propagação Direta. O valor a_6 é justamente a saída da rede neural e representa uma probabilidade do indivíduo (cujas temperatura e contagem de leucócitos foram utilizadas na computação) ter pneumonia bacteriana. Será considerado como portador da doença caso este valor seja superior a 0,5.

Os valores w e b são os parâmetros da rede neural, elementos que são aprendidos durante os treinamentos. Convém reforçar a diferença entre parâmetros e hiperparâmetros, pois esses são aprendidos pela rede e estes são definidos durante o desenvolvimento. Exemplos de hiperparâmetros são o número de camadas ocultas, a quantidade de unidades para cada camada e qual função de ativação utilizada. A escolha de hiperparâmetros adequados é fundamental para o correto funcionamento de uma rede neural. Por exemplo, caso seja escolhida a função de ativação ReLU na unidade de saída de uma rede neural com o propósito de classificação binária, o resultado final não será um valor entre 0 e 1, como uma probabilidade de pertencer a determinada classe, mas sim um valor entre 0 e infinito (36,37).

Os parâmetros são iniciados, habitualmente, de forma aleatória e o aprendizado dos melhores valores acontece após sucessivos ciclos de Propagação Direta e Propagação Reversa. A Propagação Reversa é a etapa em que as probabilidades (valor a_6) calculadas para cada indivíduo são comparadas aos diagnósticos reais, que serão representados pelo número 0, quando não há pneumonia bacteriana, e 1, quando há a doença. Esta comparação será realizada através de uma função de perda L (ou *Loss Function*). A função de erro irá retornar um valor que determina quantos erros houve durante a classificação. Um método conhecido como Descida do Gradiente (DG) irá então buscar qual ajuste deve ser feito a cada parâmetro de modo a melhorar o resultado (36). Assim, a cada ciclo, o algoritmo dá um pequeno passo em direção do melhor resultado.

Para que o aprendizado atinja resultados satisfatório, são necessários diversos ciclos de propagação direta e reversa, repassando todos os dados de treinamento diversas vezes. Dá-se o nome de época a cada vez que todos os dados são vistos pela rede neural. Oferecendo poucas épocas ao treinamento,

pode-se ocorrer o subaprendizado (ou *underfitting*), com épocas em excesso é possível incorrer no sobreaprendizado (ou *overfitting*). Estas redes artificiais são capazes de aprender tão bem os padrões dos dados de entrada, que, quando incorrem no sobreaprendizado, memorizam os rótulos dos dados apresentados, resultando em uma baixa capacidade de generalizar o seu aprendizado para dados inéditos (33,35,36).

O sobreajuste é um dos maiores desafios ao desenvolver um classificador baseado em AM, pois um sistema sobreajustado performa muito bem nos dados de treinamento, mas muito mal quando precisa ser utilizado para classificar novos pacientes. Isto pode ser exemplificado como um médico que acerta os diagnósticos somente dos casos em que ele já viu a resposta, mas não consegue reconhecer as doenças em novos pacientes. Diversas técnicas são utilizadas para evitar o sobreajuste, como um controle do tamanho da rede neural, limitação das iterações de aprendizado (*early stopping*), regularização *dropout*, entre outras (33–35,38–41).

O Aprendizado Profundo (AP) ou *Deep Learning* é um termo que muitas vezes é usado como sinônimo de Rede Neural Artificial, mas descreve, com maior precisão, aquelas redes com uma quantidade significativa de camada ocultas. Não é um consenso absoluto quantas camadas ocultas devem existir para que um sistema seja considerado profundo, mas pode se dizer que, quanto mais camadas, mais profunda é a rede. Este número pode chegar a dezenas, sobretudo em redes neurais convolutivas ou redes neurais recorrentes. Um maior número de camadas não vai, necessariamente, levar a melhores resultados, já que, com mais parâmetros, mais dados são necessários para o treinamento. Além disso, a chance de sobreajuste é maior (34,35,38–41).

RNAs são algoritmos extremamente potentes, capazes de encontrar padrões complexos e de classificar dados não linearmente separáveis. Ao mesmo tempo, exigem um grande número de dados para o treinamento, além de ter uma complexidade que aumenta exponencialmente com o número de unidades e camadas. Esta complexidade pode levar também a desafios práticos, como executar o algoritmo em máquinas com menor poder computacional (40).

2. OBJETIVOS

2.1 OBJETIVO GERAL

Analisar o desempenho de algoritmos baseados em AM como preditores de evolução durante internação hospitalar em pacientes com COVID-19, utilizando dados da admissão hospitalar.

2.2 OBJETIVOS ESPECÍFICOS

- Desenvolver formas eficientes de pré-processamento de dados clínicos, resultados de exames físico e laboratoriais, utilizando conhecimentos clínicos e matemáticos;

- Determinar o desempenho de diversos algoritmos baseados em AM e, experimentalmente, encontrar os métodos com melhor performance;

- Determinar a possibilidade e benefícios de incorporar classificadores baseados em AM para uso no cuidado de pacientes internados pela COVID-19;

- Validar uma metodologia para construção de ferramenta para auxílio no combate de novas doenças, partindo da coleta de dados (estudo de coorte), pré-processamento e experimentação de diferentes algoritmos.

3. MATERIAIS E MÉTODOS

3.1 COLETA DE DADOS

3.1.1 DESENHO DO ESTUDO

Coorte prospectiva e multicêntrica.

3.1.2 POPULAÇÃO

A população incluída neste estudo faz parte da coorte denominada “Estudo de Coorte para predição de morbiletalidade pela infecção do COVID-19: o papel dos preditores convencionais e do algoritmo baseado em Deep Learning”, desenvolvida pelos pesquisadores Dr. Hervaldo Sampaio Carvalho e Dra. Iêda Monalisa da Silva Rios. O estudo de coorte e o presente projeto foram aprovados pelo Comitê de Ética em Pesquisa da Faculdade de Medicina da Universidade de Brasília (CAAE 31941420.4.1001.5558). Os dados foram coletados mediante preenchimento do Termo de Consentimento Livre e Esclarecido pelo próprio paciente ou responsável, respeitando-se o sigilo e confidencialidade dos dados.

Pacientes acima de 18 anos internados por síndrome respiratória, por COVID-19 confirmada ou sob suspeita, em um dos hospitais incluídos no estudo: Hospital Universitário de Brasília (HUB), Brasília-DF, Brasil; Hospital Regional da Asa Norte (HRAN), Brasília-DF, Brasil; Hospital das Clínicas da Universidade Federal de Minas Gerais (UFMG), Belo Horizonte-MG, Brasil; Hospital Universitário Cassiano Antônio Moraes (HUCAM), Vitória-ES, Brasil; Hospital Universitário João de Barros Barreto da Universidade Federal do Pará (UFPA), Belém-PA, Brasil; Hospital Universitário da Universidade Federal do Vale do São Francisco (Univasf), Petrolina-PE, Brasil; Hospital Universitário da Universidade Federal da Grande Dourados (UFGD), Dourados-MS, Brasil; e Hospital de Doenças Tropicais (HDT-UFT), Araquána-TO, Brasil.

3.1.3 CRITÉRIOS DE INCLUSÃO E EXCLUSÃO

Foram incluídos no estudo pacientes acima de 18 anos internados por síndrome respiratória, por COVID-19 confirmada ou sob suspeita, em uma das unidades de saúde participantes no período de dezembro de 2019 a janeiro e 2021. Os casos de síndrome respiratória foram definidos pela associação de sintomas gerais como dispneia, queda da saturação de oxigênio, cianose e/ou alterações pulmonares radiográficas. Apenas as pacientes gestantes foram excluídas do estudo.

3.1.4 DADOS CATALOGADOS

Os dados clínicos foram coletados nas primeiras 72 horas de internação, registrando o exame clínico na admissão (incluindo sinais vitais e identificação de comorbidades), dados demográficos, exames de imagem e laboratoriais, incluindo a Reação em Cadeia da Polimerase (PCR) no escarro, *swab* nasal e a sorologia para SARS-CoV-2, na tentativa de identificar a presença do coronavírus. Após a coleta dos dados, os pacientes foram divididos em dois grupos, os que tiveram a infecção pelo SARS-CoV-2 confirmadas (seja pelo PCR ou pela sorologia) e os pacientes com exames negativos. Todos os pacientes, COVID-19 positivos ou não, também foram divididos em pacientes que apresentaram complicações graves e pacientes que não apresentaram estas complicações. Foram considerados graves os pacientes com indicação de ventilação mecânica, intubação orotraqueal ou acesso à Unidade de Terapia Intensiva (UTI), os que apresentaram parada cardiorrespiratória e/ou sepse grave e os que vieram a óbito. A **Tabela 2** descreve em detalhes os dados coletados.

Tabela 2. Dados coletados para cada participante do estudo

DADOS DEMOGRÁFICOS	Nome
	Data da Internação
	Endereço, telefone, e-mail
	Data de Nascimento e idade
	Cor
	Sexo
	Estado Civil
	Escolaridade
	Renda Familiar
	Moradia (tipo, número de cômodos e de moradores)
ANAMNESE E EXAME FÍSICO NA ADMISSÃO	Medicações em uso prévio
	Doenças prévias
	Sintomas possivelmente relacionados COVID-19 e outros sintomas
	Data de início dos sintomas de COVID-19
	Queixa de dificuldade respiratória
	Estado Geral
	Pressão Arterial, Frequência Cardíaca e Respiratória
	Temperatura
	Saturação de oxigênio (SpO ₂)
	Altura, Peso e Índice de Massa Corporal (IMC)
	Alterações no exame físico nas primeiras 24 horas da admissão
	Sinais de Insuficiência Respiratória nas primeiras 24 horas
	Abertura Ocular, Resposta Verbal e Motora, Escala de Glasgow
	Terapêutica para SARS-CoV-2 ou complicações
EXAMES LABORATORIAIS	PCR do Swab nasal e lavado traqueal
	Sorologia para SARS-CoV-2
	Hemograma Completo
	Velocidade de hemossedimentação (VHS)
	Uréia e Creatinina

	Sódio, Potássio, Cálcio e Magnésio
	Glicose e Hemoglobina Glicosilada
	Ácido Úrico
	Albumina
	Colesterol total e frações
	Triglicerídeos
	Desidrogenase Láctica
	Creatinina Quinase ou Fosfoquinase (CK ou CPK)
	Enzima Aminotransferase de Aspartato e de Alanina (TGO e TGO)
	Gama GT
	Bilirrubina Total e frações
	Tempo de Tromboplasmina Parcial Ativado (PTTa)
	Tempo de protrombina (TAP)
	Fosfatase Alcalina
	Transferrina e Ferritina
	Proteína C Reativa (PCR)
	D-Dímero
	Troponina T
	Fibrinogênio
	Cortisol sérico entre 7 e 9h
	TSH e T4L
	Proteinúria
EXAMES DE IMAGEM	Radiografia de Tórax Posteroanterior, Perfil e Anteroposterior
	Tomografia Computadorizada de Tórax
	Tomografia Computadorizada de Tórax para Score de Cálcio
EVOLUÇÃO	Presença de Indicação e/ou acesso à ventilação mecânica
	Intubação Orotraqueal
	Tempo de intubação e de ventilação mecânica
	Presença de Indicação e/ou acesso à Unidade de Terapia Intensiva
	Parada Cardiorrespiratória abortada

Sepse grave
Óbito
Data da Alta Hospitalar
Necessidade de reinternação

3.2 PRÉ-PROCESSAMENTO DOS DADOS

Para o desenvolvimento de algoritmos preditores de gravidade os dados passaram por um pré-processamento. Diante da escassez de recursos durante a pandemia, uma série de exames não pôde ser realizada em todos os pacientes, assim, há um elevado número de dados faltantes entre os indivíduos. Dados faltantes ou incompletude são um problema em diversos estudos realizados durante a pandemia (18). Inicialmente, foram eliminadas as variáveis coletadas em menos de 50% dos participantes e as variáveis constantes, que apresentaram o mesmo valor para todos os participantes, como por exemplo a presença de turgência jugular, que foi negativa em todos os pacientes.

Ao trabalhar com dados incompletos, dois caminhos podem ser percorridos. O primeiro consiste em eliminar todos os indivíduos que não tiveram as variáveis de interesse para o algoritmo preenchidas. O segundo é realizar imputação de dados, ou seja, completar os dados faltantes com estimativas e, assim, manter todos os indivíduos no estudo (18). Há diversas maneiras de realizar imputação, neste estudo todas as variáveis foram imputadas utilizando a média. A título de exemplo, o valor médio da hemoglobina de todos os pacientes foi utilizado para preencher a dosagem da hemoglobina daqueles em que o exame não foi realizado. Ambas as abordagens foram utilizadas neste trabalho.

Além dos dados faltantes, as condições clínicas que não foram observadas em um número significativos de sujeitos também são desafios encontrados no desenvolvimento de preditores que utilizam dados oriundos de estudos clínicos (42). Por estarem positivas em uma parcela muito pequena de indivíduos, estas variáveis clínicas não trazem muitas informações para os

classificadores e aumentam a sua complexidade, ao aumentar a dimensionalidade dos dados. Uma série de variáveis coletadas em nossa coorte, sobretudo doenças prévias, sintomas ou alterações ao exame físico, foram positivas em menos de 1% dos pacientes. Na tentativa de reduzir a dimensionalidade dos dados, sem eliminar a importância dessas alterações, foram catalogadas as variáveis passíveis de agrupamento ou clusterização utilizando o conhecimento médico. Assim, variáveis com significado clínico semelhantes ou que compõem uma mesma síndrome foram unidas através de uma operação lógica "OU", de modo que a variável que representa o agrupamento seria negativa apenas se todas as que a originaram também forem negativas. Este procedimento está exemplificado na **Tabela 3**.

Tabela 3. Exemplo da operação lógica "OU" para agrupamento de variáveis. Suponha que as variáveis de A a D possuem significado clínico semelhante ou constituem mesma síndrome, a coluna "agrupamento" representa a união destas variáveis

	Variável A	Variável B	Variável C	Variável D	Agrupamento
Paciente 1	Negativo	Positivo	Negativo	Negativo	Positivo
Paciente 2	Positivo	Negativo	Positivo	Positivo	Positivo
Paciente 3	Negativo	Negativo	Negativo	Negativo	Negativo
Paciente 4	Negativo	Negativo	Negativo	Positivo	Positivo

Com este agrupamento ou clusterização, criou-se uma variável para a presença de doenças pulmonares prévias, unindo a presença de Doença Pulmonar Obstrutiva Crônica (DPOC), Asma ou outra doença pulmonar crônica. O mesmo foi realizado para sintomas de Infecção de Vias Aéreas Superiores (IVAS), sintomas inespecíficos de infecção viral e sinais de acometimento pulmonar, compondo outros 3 grandes grupos. Variáveis com baixa representação, ou seja, menos de 5% de casos positivos, mesmo após agrupamentos, foram eliminadas da análise. Variáveis cujo significado está ligado a alguma outra alteração, como por exemplo perda do paladar e anosmia, também foram eliminadas, preferindo aquela com maior frequência de casos

positivos. A distribuição das variáveis para cada classe (ou desfecho) não foi considerada, de modo a evitar viés na análise dos classificadores.

As variáveis numéricas foram categorizadas. A idade foi dividida em grupos baseados em percentis, utilizando 0, 10%, 25%, 50%, 75%, 90% e 100% como separadores. Assim, pacientes cuja idade se encontram iguais ou abaixo do percentil 10% receberão a categoria “0” para a idade. Receberam o valor “1” aqueles cuja idade está acima do percentil 10%, mas igual ou inferior ao 25%, “2” quando entre 25% e 50% e assim por diante. Esta abordagem faz com que a idade deixe de ser representada em anos e passe a ser representada por uma classe que vai de 0 a 5.

Já a hemoglobina foi categorizada, nas pacientes do sexo feminino, utilizando os intervalos [0, 6.5], (6.5, 8], (8, 10], (10, 11.5], (11.5, 16.5] e superior a 16,5g/dL. Nos do sexo masculino foram utilizados os intervalos [0, 6.5], (6.5, 8], (8, 10], (10, 13], (13, 18] e superior a 18g/dL. Deste modo, uma mulher com hemoglobina de 8g/dL recebeu a categoria “1” e um homem com hemoglobina de 13,7g/dL recebeu a categoria “4”. Esta abordagem foi inspirada na classificação de anemias pelo *National Health Service* (NSH) e *National Cancer Institute* (NCI) para classificação de anemias e suas gravidades (43,44).

Leucócitos, linfócitos, plaquetas e ureia foram categorizados em reduzidos, normais e elevados, também inspirados nos valores de referência definidos pelo *National Health Service* (45). A **Tabela 4** define os valores de referência utilizados e a categoria atribuída.

Tabela 4. Categorização dos leucócitos, linfócitos, plaquetas e ureia

	“0” - Reduzido	“1” - Normal	“2” - Elevado
Leucócitos (número/ μ L)	≤ 4.000	(4.000, 11.000]	> 11.000
Linfócitos (número/ mm^3)	≤ 1	(1, 4]	> 4
Plaquetas (número/ μ L)	≤ 150.000	(150.000, 400.000]	> 400.000
Ureia (mg/dL)	$\leq 15,02$	(15,02, 46,847]	$> 46,847$

Apesar de terem sido aferidas à admissão do paciente, as pressões arteriais sistólica e diastólica foram categorizadas utilizando os critérios inspirados na classificação da American Heart Association (AHA) (45) para hipertensão e hipotensão, eliminando apenas a faixa entre 140 e 180 da pressão arterial sistólica (pois havia somente 2 casos acima de 180mmHg). Este modo de classificação da pressão arterial foi escolhido apenas para utilizar os valores como referência e não para diagnosticar e classificar hipertensão arterial. Assim, as faixas para pressão sistólica foram [0, 90], (90, 120], (120, 130], (130, 140] e superior a 140mmHg. Para a diastólica, [0, 60], (60, 80], (80, 90], (90, 120] e superior a 120mmHg.

A frequência cardíaca foi categorizada utilizando 60bpm como valor de corte para bradicardia e 100bpm como valor de corte para taquicardia. Já a frequência respiratória foi definida como normal até 16irpm, limítrofe entre 16 e 20, elevada entre 20 e 24 e muito elevada acima de 24. Por fim, valores iguais ou inferiores a 88% para a Saturação Periférica de O₂ (SpO₂) foram considerados como extremamente baixos, entre 88% e 92% inclusive como muito baixos, entre 92% e 96% inclusive como baixos e os acima de 96% como normal.

Esta categorização foi realizada de modo a aproximar o valor numérico entre as diferentes variáveis, já que a hemoglobina variava em torno do número 10 enquanto as plaquetas variavam em torno de 200.000. Esta diferença na faixa dinâmica das variáveis e na ordem de grandeza poderia resultar em pior performance, já que certos algoritmos poderiam atribuir maior importância a determinadas variáveis não pela sua importância na predição, mas pelo seu valor absoluto. Outra abordagem que poderia ser utilizada é a da normalização, ao deixar todas as variáveis numéricas entre 0 e 1. Isto não foi realizado pois a categorização também aproximou as variáveis numéricas (como resultados de exames laboratoriais) das categóricas (como por exemplo o estado geral, que foi classificado em 0: bom estado geral, 1: levemente comprometido, 2: moderadamente comprometido, 3: gravemente comprometido e 4: crítico).

3.3 CONSTRUÇÃO DOS CONJUNTOS DE DADOS

De modo a verificar o impacto que a eliminação de indivíduos com dados faltantes e a imputação poderia gerar no treinamento dos algoritmos, foram selecionadas as 28 variáveis mais completas, ou seja, as que foram coletadas na maior parte dos pacientes. Com estas variáveis, 7 conjuntos de dados (*datasets*) diferentes foram construídos (conjunto A ao G), conforme ilustrado na **Tabela 5**, que também indica quantos pacientes foram incluídos em cada conjunto. Um conjunto (G) incluiu todos os 421 pacientes e todas as 28 variáveis, que foram completadas utilizando a imputação pela média. Nos demais (conjuntos A ao F) não foi realizada imputação, resultando em perda de indivíduos pela eliminação dos dados incompletos. Estes 6 conjuntos foram construídos variando o número de variáveis que o compõem, indo das 23 variáveis mais completas até um conjunto com todas as 28 variáveis.

Tabela 5. Variáveis, imputação e quantidade de indivíduos em cada conjunto de dados

Conjunto A (23 variáveis, 355 pacientes)	Conjunto B (24 variáveis, 335 pacientes)	Conjunto C (25 variáveis, 309 pacientes)	Conjunto D (26 variáveis, 285 pacientes)	Conjunto E (27 variáveis, 285 pacientes)	Conjunto F (28 variáveis, 246 pacientes)	Conjunto G (28 variáveis com imputação, 421 pacientes)
Idade						
Hemoglobina						
Leucócitos						
Linfócitos						
Plaquetas						
Frequência cardíaca na admissão						
Saturação de oxigênio na Admissão						
Ausência de doença prévia						
Hipertensão arterial sistêmica prévia						
Diabetes Mellitus prévio						
Obesidade						
Insuficiência renal						
Febre						
Tosse Seca						
Cefaleia						
Fraqueza						
Anosmia						
Sexo						
Sinais de IVAS						
Dispneia						
Sinais de acometimento pulmonar						
Sinais inespecíficos de infecção viral						
Doenças pulmonares prévias						
Estado geral na admissão						
Frequência respiratória na admissão						
Pressão arterial sistólica na admissão						
Pressão arterial diastólica na admissão						
Ureia						

3.4 DESENVOLVIMENTO DO CLASSIFICADOR

Os dados foram então separados em grupos de treinamento e teste, respectivamente com 80% e 20% do total de pacientes, e utilizados para treinar algoritmos de *Machine Learning* (ML). O grupo de treinamento é o que será apresentado ao algoritmo em conjunto com o desfecho observado pela coorte, enquanto o de teste é apresentado somente após o treinamento, sem o desfecho real, para avaliação da performance. Esta seleção aleatória entre os grupos pode

introduzir viés na análise da acurácia, a exemplo de quando indivíduos mais fáceis de classificar são agrupados no grupo de teste, aumentando a acurácia do sistema. Para minimizar este impacto, cada algoritmo foi treinado em 5000 simulações, com reamostragens aleatórias dos grupos. Esta técnica de validação com diferentes reamostragens dos grupos de treino e teste é conhecida como Validação Cruzada de Monte Carlo e é mais precisa mesmo em conjuntos de dados com mais de 1000 observações (46). Durante cada simulação foram calculadas a acurácia, sensibilidade, especificidade, valor preditivo negativo, valor preditivo positivo, F1-Score e Área Sob a Curva (AUC) Característica de Operação do Operador (ROC). Avaliou-se então os valores mínimos e máximos, média e desvio padrão de cada métrica descrita. As bibliotecas utilizadas para o treinamento dos algoritmos foram *Lazy Predict* 0.2.12, *Scikit-Learn* 1.2.0 (47) e *Tensorflow* 2.10.1 (48), executados em *Python* 3.9.

Os testes com os diferentes algoritmos foram realizados da seguinte forma:

1. 100 simulações utilizando a função *LazyClassifier*, da biblioteca *Lazy Predict* 0.2.12, que automatiza o treinamento de 27 algoritmos diferentes baseados em ML. São eles:
 - a. *Stochastic Gradient Descent Classifier*;
 - b. *Linear Support Vector Classification*;
 - c. *Logistic Regression*;
 - d. *Linear Discriminant Analysis*;
 - e. *Ridge Classifier with Cross Validation*;
 - f. *Ridge Classifier*;
 - g. *AdaBoost Classifier*;
 - h. *Nearest Centroid*;
 - i. *Gaussian Naive Bayes*;
 - j. *Passive Aggressive Classifier*;
 - k. *Nu Support Vector Classification*;
 - l. *Light Gradient Boosting Machine Classifier*;
 - m. *Quadratic Discriminant Analysis*;
 - n. *Support Vector Classification*;

- o. *XGBoost Classifier*;
 - p. *Bernoulli Naive Bayes*;
 - q. *KNeighbors Classifier*;
 - r. *Random Forest Classifier*;
 - s. *Bagging Classifier*;
 - t. *Perceptron*;
 - u. *Extra Trees Classifier*;
 - v. *Extra Tree Classifier*;
 - w. *Calibrated Classifier with Cross Validation*;
 - x. *Label Spreading*;
 - y. *Label Propagation*;
 - z. *Decision Tree Classifier*;
 - aa. *Dummy Classifier*;
2. Análise da acurácia média para cada um dos 27 algoritmos testados, para cada um dos 7 conjuntos de dados construídos neste trabalho;
 3. Para cada conjunto de dados, foi selecionado o algoritmo com melhor desempenho e sua acurácia média foi registrada;
 4. Partindo da maior acurácia média atingida por cada conjunto de dados, foi selecionado o conjunto que apresentou melhor desempenho;
 5. Seleção dos 7 algoritmos com maior acurácia média no conjunto de dados selecionado, para retreinamento em 5000 simulações com diferentes composições de grupo de treinamento e teste (Validação Cruzada de Monte Carlo);
 - a. Estes algoritmos foram treinados utilizando a biblioteca *Scikit-Learn 1.2.0*
 6. Este conjunto de dados selecionado na etapa 4 também foi utilizado para o treinamento de redes neurais densas, com diferentes números de camadas e de unidades por camada. Cada rede neural foi treinada em 100 simulações para determinar aquela com maior acurácia média;
 - a. A rede neural foi construída, treinada e avaliada utilizando a biblioteca *Tensorflow 2.10.1*

- b. A **Figura 10** exemplifica uma configuração de rede neural densa, capaz de receber 28 variáveis na entrada e com 2 camadas escondidas, com 8 e 4 unidades em cada. A cada opção, como número de camadas e unidades por camada, da-se o nome de hiperparâmetros.
7. A rede neural densa com maior acurácia média foi então selecionada para mais 5000 treinamentos, também utilizando a Validação Cruzada de Monte Carlo para avaliar seu desempenho.

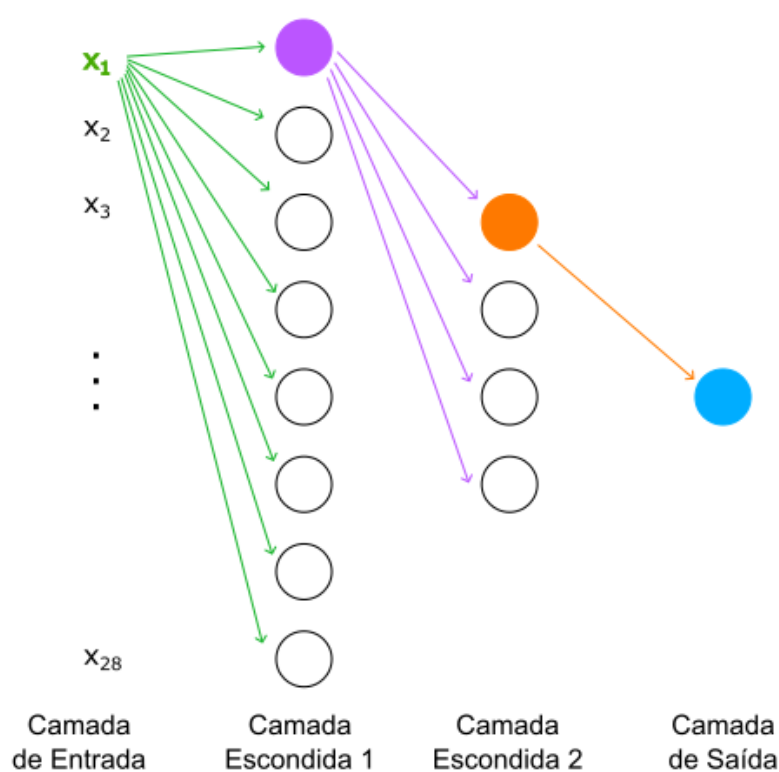


Figura 10. Exemplo de rede neural densa, recebendo 28 variáveis na entrada e constituída de 8 unidades na primeira camada, 4 unidades na segunda e 1 unidade de saída. Apenas as conexões da primeira variável e das primeiras unidades de cada camada foram representadas de modo a simplificar a imagem. Fonte: Autor

As redes neurais utilizadas neste trabalho diferem entre si pelo número de camadas, número de unidades por camada e pela taxa de *Dropout*. O *Dropout* é uma técnica de regularização que consiste em, durante o treinamento da rede neural, ignorar uma porcentagem das unidades em uma camada (quantidade determinada pela taxa de *Dropout*). Ao desconsiderar parte das unidades, de

forma aleatória, esta técnica distribui o aprendizado entre as conexões e auxilia a reduzir o *overfitting* ou superaprendizado, melhorando a generalização do modelo e o desempenho da rede neural ao receber dados inéditos (39).

4. RESULTADOS

4.1 ANÁLISE DE DADOS DA COORTE: DESFECHOS

Dos 537 participantes do estudo, 421 tiveram o diagnóstico de COVID-19 confirmado. Os 116 que não tiveram a infecção documentada foram excluídos do estudo. Destes 421 com doença confirmada, 133 tiveram um desfecho considerado grave (31,5%). A **Figura 11** apresenta o número de pacientes com cada desfecho grave. É importante ressaltar que um mesmo paciente pode apresentar 2 ou mais desfechos graves durante o período de acompanhamento, o que pode ser visualizado nesta figura pelo preenchimento de diferentes cores em uma mesma coluna.

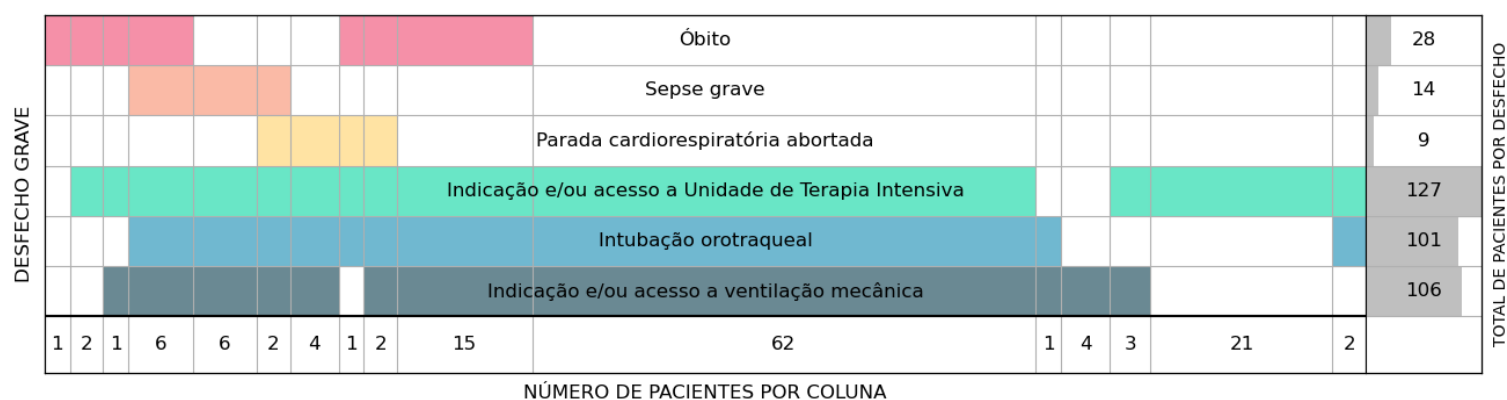


Figura 11. Número de pacientes para cada desfecho grave. Colunas preenchida por diferentes cores representam casos de ocorrência de diferentes desfechos no mesmo paciente. Por exemplo, à coluna central observou-se que 62 pacientes apresentaram indicação de acesso à UTI, de intubação orotraqueal e de ventilação mecânica, mas não os demais desfechos.

4.2 ANÁLISE DE DADOS DA COORTE: ANÁLISE DEMOGRÁFICA

O perfil demográfico dos grupos de pacientes que evoluíram com forma grave (n = 133) e não grave (n = 288) foi analisado, conforme descrito nas **Tabelas 6 a 11**. Pode-se observar que há um predomínio de pacientes do sexo feminino, de etnia parda e com idade superior a 50 anos de idade.

Tabela 6. Distribuição dos gêneros

Gênero	Não grave (risco absoluto)	Grave (risco absoluto)	Total
Masculino	128 (71,91%)	50 (28,09%)	178
Feminino	160 (65,84%)	83 (34,16%)	243

Tabela 7. Distribuição das etnias

Etnia	Não grave (risco absoluto)	Grave (risco absoluto)	Total
Branca	53 (68,83%)	24 (31,17%)	77
Preta	17 (73,91%)	6 (26,09%)	23
Parda	189 (66,55%)	95 (33,45%)	284
Amarela	5 (62,50%)	3 (37,50%)	8
Indígena	2 (66,67%)	1 (33,33%)	3
Sem declaração	19 (86,36%)	3 (13,64%)	22
Não preenchido	3 (75,00%)	1 (25,00%)	4

Tabela 8. Distribuição do estado civil

Estado Civil	Não grave (risco absoluto)	Grave (risco absoluto)	Total
Solteiro (a)	72 (69,23%)	32 (30,77%)	104
Casado (a)	150 (73,17%)	55 (26,83%)	205
Divorciado (a)	26 (72,22%)	10 (27,78%)	36
Viúvo (a)	12 (57,14%)	9 (42,86%)	21
Outro	24 (52,17%)	22 (47,83%)	46
Não preenchido	4 (44,44%)	5 (55,56%)	9

Tabela 9. Distribuição da renda familiar

Renda Familiar	Não grave	Grave	Total
Até 1 salário-mínimo	33 (73,33%)	12 (26,67%)	45
1 a menos de 2 salários-mínimos	45 (70,31%)	19 (29,69%)	64
2 a menos de 3 salários-mínimos	37 (78,72%)	10 (21,28%)	47
3 a menos de 5 salários-mínimos	25 (73,53%)	9 (26,47%)	34
5 a menos de 10 salários-mínimos	17 (85,00%)	3 (15,00%)	20
10 ou mais salários-mínimos	10 (66,67%)	5 (33,33%)	15
Sem rendimento	3 (75,00%)	1 (25,00%)	4
Não preenchido	118 (61,46%)	74 (38,54%)	192

Tabela 10. Distribuição do tipo de moradia

Tipo de Moradia	Não grave	Grave	Total
Própria	130 (72,63%)	49 (27,37%)	179
Alugada	28 (84,85%)	5 (15,15%)	33
Cedida	9 (64,29%)	5 (35,71%)	14
Não preenchido	121 (62,05%)	74 (37,95%)	195

Tabela 11. Análise das variáveis demográficas numéricas: idade do paciente, número de cômodos na casa que habita e número de moradores que vivem na mesma casa.

	Média			Mediana			Desvio Padrão			Não preenchidos		
	Total	Não Grave	Grave	Total	Não Grave	Grave	Total	Não Grave	Grave	Total	Não Grave	Grave
Idade (anos)	57,57	55,60	61,82	57,00	55,00	62,00	15,98	15,35	16,54	1	1	0
Número de cômodos	6,42	6,29	7,05	5,00	5,00	7,00	2,36	2,29	2,67	308	195	113
Número de Moradores	3,41	3,47	3,19	3,00	3,00	3,00	1,65	1,66	1,62	261	164	97

4.3 ANÁLISE GLOBAL DAS VARIÁVEIS UTILIZADAS NO ESTUDO

A **Tabela 12** demonstra as médias e desvios padrão para cada grupo (grave e não grave) e o número de dados faltantes (campos que não foram preenchidos durante a coleta de dados) para cada variável numérica. Apenas aquelas com mais de 50% de respostas foram consideradas.

A **Tabela 13** descreve as variáveis categóricas coletadas à admissão, separadas entre os grupos grave e não grave. As **Tabelas 14 e 15** descrevem as distribuições para dispneia e estado geral do paciente à admissão. Se destacam o número de variáveis categóricas que praticamente não foram identificadas durante a coorte, à exemplo da arritmia cardíaca e doença valvular, que estiveram presentes em somente um dos 421 pacientes. Isto faz com que os dados sejam muito esparsos e gera certas dificuldades ao desenvolver algoritmos de classificação que vão além do AM, que serão discutidos a seguir.

Tabela 12. Análise das variáveis numéricas com relação ao desfecho.

Variável Numérica	Não grave		Grave		Dados faltantes
	Média	Desvio Padrão	Média	Desvio Padrão	
Idade (anos)	55,60	15,35	61,82	16,54	1
Hemoglobina (g/dL)	12,63	2,23	12,16	2,50	5
Hematócrito (%)	38,79	6,79	36,53	7,47	97
Número de Leucócitos (número/mcL)	6185,40	9562,86	6892,14	5950,38	14
Número de Linfócitos (número/mm ³)	532,57	701,50	463,44	578,70	18
Número de Plaquetas (número/ μ L)	152531,25	155866,64	122447,66	142226,82	7
Ureia (mg/dL)	45,62	33,30	59,20	44,98	56
Creatinina (mg/dL)	1,31	1,53	1,69	2,21	80
Sódio (Na) (mg/dL)	138,58	22,08	138,77	19,64	169
Potássio (K) (mg/dL)	4,31	1,58	4,53	1,96	171
Glicose (mg/dL)	157,05	97,65	206,61	118,20	162
Desidrogenase Láctica - DHL (mg/dL)	2471,32	22271,52	1454,03	7527,05	199
Enzima Aminotransferase de Aspartate (TGO) (mg/dL)	85,36	343,58	93,53	322,12	110
Enzima Aminotransferase de Alanine (TGP) (mg/dL)	96,61	350,61	77,75	216,22	127
Proteína C Reativa (PCR) (mg/dL)	51,10	59,48	77,56	85,60	206
Pressão Arterial Sistólica na admissão (mmHg)	124,67	20,13	125,43	20,53	54
Pressão Arterial Diastólica na admissão (mmHg)	78,54	14,09	76,15	13,48	54
Frequência Cardíaca na admissão (bpm/min)	86,22	15,99	88,84	17,47	29
Frequência Respiratória na admissão (irpm/min)	22,48	4,94	24,56	6,83	59
Saturação de oxigênio na admissão (%)	94,24	3,73	91,17	6,91	35
Peso na admissão (Kg)	82,32	19,09	78,35	18,51	178

Tabela 13. Variáveis categóricas.

Variável	Negativo			Positivo				
	Não Grave	Grave	Total	Não Grave	Grave	Total		
Doença prévia	Nenhuma doença prévia	219 (64,41%)	121 (35,59%)	340	69 (85,19%)	12 (14,81%)	81	
	Hipertensão Arterial	144 (71,29%)	58 (28,71%)	202	144 (65,75%)	75 (34,25%)	219	
	Doença Isquêmica do Coração	279 (69,06%)	125 (30,94%)	404	9 (52,94%)	8 (47,06%)	17	
	Insuficiência Cardíaca	284 (69,27%)	126 (30,73%)	410	4 (36,36%)	7 (63,64%)	11	
	Arritmia cardíaca	288 (68,57%)	132 (31,43%)	420	0 (0,00%)	1 (100,00%)	1	
	Doença Valvular cardíaca	287 (68,33%)	133 (31,67%)	420	1 (100,00%)	0 (0,00%)	1	
	Dislipidemia	281 (69,04%)	126 (30,96%)	407	7 (50,00%)	7 (50,00%)	14	
	Diabetes Mellitus	206 (71,28%)	83 (28,72%)	289	82 (62,12%)	50 (37,88%)	132	
	Obesidade	259 (68,70%)	118 (31,30%)	377	29 (65,91%)	15 (34,09%)	44	
	Insuficiência Renal	270 (68,88%)	122 (31,12%)	392	18 (62,07%)	11 (37,93%)	29	
	DPOC	275 (68,41%)	127 (31,59%)	402	13 (68,42%)	6 (31,58%)	19	
	Asma brônquica	279 (68,05%)	131 (31,95%)	410	9 (81,82%)	2 (18,18%)	11	
	Outra Doença pulmonar crônica	286 (68,59%)	131 (31,41%)	417	2 (50,00%)	2 (50,00%)	4	
	Câncer maligno de qualquer origem	279 (68,38%)	129 (31,62%)	408	9 (69,23%)	4 (30,77%)	13	
	Artrite Reumatoide	287 (68,33%)	133 (31,67%)	420	1 (100,00%)	0 (0,00%)	1	
	Lúpus Eritematoso sistêmico	288 (68,57%)	132 (31,43%)	420	0 (0,00%)	1 (100,00%)	1	
	Outra doença reumática	286 (68,26%)	133 (31,74%)	419	2 (100,00%)	0 (0,00%)	2	
	Doença hepática crônica	284 (68,27%)	132 (31,73%)	416	4 (80,00%)	1 (20,00%)	5	
	Doenças gastrointestinais crônicas	286 (68,42%)	132 (31,58%)	418	2 (66,67%)	1 (33,33%)	3	
	Acidente Vascular Cerebral	287 (68,66%)	131 (31,34%)	418	1 (33,33%)	2 (66,67%)	3	
	Síndrome demencial	287 (68,50%)	132 (31,50%)	419	1 (50,00%)	1 (50,00%)	2	
	Outra doença neurológica crônica	283 (68,19%)	132 (31,81%)	415	5 (83,33%)	1 (16,67%)	6	
	Doença genética	287 (68,33%)	133 (31,67%)	420	1 (100,00%)	0 (0,00%)	1	
	Portador de necessidade especial	288 (68,57%)	132 (31,43%)	420	0 (0,00%)	1 (100,00%)	1	
	Outra(s) doença(s)	221 (67,17%)	108 (32,83%)	329	67 (72,83%)	25 (27,17%)	92	
	Presença de sintoma	Nenhum sintoma	285 (68,67%)	130 (31,33%)	415	3 (50,00%)	3 (50,00%)	6
		Febre	145 (66,82%)	72 (33,18%)	217	143 (70,10%)	61 (29,90%)	204
Tosse seca		150 (64,94%)	81 (35,06%)	231	138 (72,63%)	52 (27,37%)	190	
Tosse produtiva		270 (68,88%)	122 (31,12%)	392	18 (62,07%)	11 (37,93%)	29	
Coriza,		265 (69,37%)	117 (30,63%)	382	23 (58,97%)	16 (41,03%)	39	
Corrimento nasal posterior		284 (68,11%)	133 (31,89%)	417	4 (100,00%)	0 (0,00%)	4	
Congestão nasal		281 (68,37%)	130 (31,63%)	411	7 (70,00%)	3 (30,00%)	10	
Espirro		285 (68,51%)	131 (31,49%)	416	3 (60,00%)	2 (40,00%)	5	
Dor de garganta		276 (69,00%)	124 (31,00%)	400	12 (57,14%)	9 (42,86%)	21	
Dor torácica		269 (68,80%)	122 (31,20%)	391	19 (63,33%)	11 (36,67%)	30	

Variável	Negativo		Total	Positivo		Total	
	Não Grave	Grave		Não Grave	Grave		
Dor de cabeça	216 (64,29%)	120 (35,71%)	336	72 (84,71%)	13 (15,29%)	85	
Tontura	280 (68,46%)	129 (31,54%)	409	8 (66,67%)	4 (33,33%)	12	
Turvação visual	287 (68,50%)	132 (31,50%)	419	1 (50,00%)	1 (50,00%)	2	
Fraqueza	212 (67,09%)	104 (32,91%)	316	76 (72,38%)	29 (27,62%)	105	
Dor muscular	228 (66,67%)	114 (33,33%)	342	60 (75,95%)	19 (24,05%)	79	
Dor articular	276 (67,98%)	130 (32,02%)	406	12 (80,00%)	3 (20,00%)	15	
Lesões de pele	288 (68,74%)	131 (31,26%)	419	0 (0,00%)	2 (100,00%)	2	
Anosmia	258 (68,07%)	121 (31,93%)	379	30 (71,43%)	12 (28,57%)	42	
Redução paladar (disgeusia)	266 (68,38%)	123 (31,62%)	389	22 (68,75%)	10 (31,25%)	32	
Diarreia	254 (66,84%)	126 (33,16%)	380	34 (82,93%)	7 (17,07%)	41	
Dor abdominal	279 (68,22%)	130 (31,78%)	409	9 (75,00%)	3 (25,00%)	12	
Confusão mental	284 (69,61%)	124 (30,39%)	408	4 (30,77%)	9 (69,23%)	13	
Calafrios	270 (68,18%)	126 (31,82%)	396	18 (72,00%)	7 (28,00%)	25	
Caroços embaixo da pele (gânglios aumentados)	286 (68,26%)	133 (31,74%)	419	2 (100,00%)	0 (0,00%)	2	
Náusea	270 (67,84%)	128 (32,16%)	398	18 (78,26%)	5 (21,74%)	23	
Vômito	271 (68,09%)	127 (31,91%)	398	17 (73,91%)	6 (26,09%)	23	
Inapetência	257 (67,81%)	122 (32,19%)	379	31 (73,81%)	11 (26,19%)	42	
Boca seca	286 (68,42%)	132 (31,58%)	418	2 (66,67%)	1 (33,33%)	3	
Baixo débito urinário	280 (69,14%)	125 (30,86%)	405	8 (50,00%)	8 (50,00%)	16	
Dificuldade para deglutir	287 (68,66%)	131 (31,34%)	418	1 (33,33%)	2 (66,67%)	3	
Extremidades roxas (cianose)	288 (69,23%)	128 (30,77%)	416	0 (0,00%)	5 (100,00%)	5	
Outro(s) sintoma(s)	249 (67,12%)	122 (32,88%)	371	39 (78,00%)	11 (22,00%)	50	
Alteração ao exame físico nas primeiras 24h	Ausência de alterações ao exame físico	142 (60,68%)	92 (39,32%)	234	146 (78,07%)	41 (21,93%)	187
	Hiperemia conjuntival	287 (68,50%)	132 (31,50%)	419	1 (50,00%)	1 (50,00%)	2
	Mucosas descoradas	288 (69,23%)	128 (30,77%)	416	0 (0,00%)	5 (100,00%)	5
	Secreção nasal	281 (68,54%)	129 (31,46%)	410	7 (63,64%)	4 (36,36%)	11
	Corrimento posterior	287 (68,66%)	131 (31,34%)	418	1 (33,33%)	2 (66,67%)	3
	Presença de anosmia	279 (67,72%)	133 (32,28%)	412	9 (100,00%)	0 (0,00%)	9
	Presença de dor a compressão de seios da face	287 (68,33%)	133 (31,67%)	420	1 (100,00%)	0 (0,00%)	1
	Presença de dor à compressão torácica	283 (68,86%)	128 (31,14%)	411	5 (50,00%)	5 (50,00%)	10
	Presença de ritmo cardíaco irregular	288 (68,57%)	132 (31,43%)	420	0 (0,00%)	1 (100,00%)	1
	Presença de hepatomegalia	287 (68,33%)	133 (31,67%)	420	1 (100,00%)	0 (0,00%)	1
	Presença de dor à palpação abdominal	280 (67,96%)	132 (32,04%)	412	8 (88,89%)	1 (11,11%)	9

Variável	Negativo			Positivo		
	Não Grave	Grave	Total	Não Grave	Grave	Total
Presença de dor articular	286 (68,42%)	132 (31,58%)	418	2 (66,67%)	1 (33,33%)	3
Presença de inflamação articular	286 (68,42%)	132 (31,58%)	418	2 (66,67%)	1 (33,33%)	3
Presença de edema de MMII	274 (68,50%)	126 (31,50%)	400	14 (66,67%)	7 (33,33%)	21
Presença de dor à palpação muscular	285 (68,51%)	131 (31,49%)	416	3 (60,00%)	2 (40,00%)	5
Presença lesões de pele	284 (68,60%)	130 (31,40%)	414	4 (57,14%)	3 (42,86%)	7
Presença de adenomegalia	287 (68,33%)	133 (31,67%)	420	1 (100,00%)	0 (0,00%)	1
Outras alterações	278 (69,50%)	122 (30,50%)	400	10 (47,62%)	11 (52,38%)	21
Sinais de Insuficiência Respiratória nas primeiras 24h						
Esforço respiratório (tiragem, cornagem, batimento de asa do nariz)	286 (69,25%)	127 (30,75%)	413	2 (25,00%)	6 (75,00%)	8
Cianose central	288 (68,57%)	132 (31,43%)	420	0 (0,00%)	1 (100,00%)	1
Aumento da frequência respiratória	280 (70,71%)	116 (29,29%)	396	8 (32,00%)	17 (68,00%)	25
Tipo respiratório irregular (Biot, Kusmaull, Cheyne-Stokes)	286 (68,42%)	132 (31,58%)	418	2 (66,67%)	1 (33,33%)	3
Diminuição da expansibilidade torácica	287 (68,33%)	133 (31,67%)	420	1 (100,00%)	0 (0,00%)	1
Presença de roncos	285 (68,51%)	131 (31,49%)	416	3 (60,00%)	2 (40,00%)	5
Presença de sibilos	287 (68,50%)	132 (31,50%)	419	1 (50,00%)	1 (50,00%)	2
Presença de estertores crepitantes	265 (68,65%)	121 (31,35%)	386	23 (65,71%)	12 (34,29%)	35
Presença de expectoração rósea e sinais de edema pulmonar	288 (68,57%)	132 (31,43%)	420	0 (0,00%)	1 (100,00%)	1
Outros sinais	286 (68,59%)	131 (31,41%)	417	2 (50,00%)	2 (50,00%)	4
Uso de terapêutica específica para SARS-CoV-2 ou complicações						
Nenhuma Terapêutica específica	276 (68,32%)	128 (31,68%)	404	12 (70,59%)	5 (29,41%)	17
Hidroxicloroquina e Cloroquina	271 (68,26%)	126 (31,74%)	397	17 (70,83%)	7 (29,17%)	24
Nitazoxanida (Anita)	280 (67,80%)	133 (32,20%)	413	8 (100,00%)	0 (0,00%)	8
Azitromicina	119 (72,12%)	46 (27,88%)	165	169 (66,02%)	87 (33,98%)	256
Remdesivir	288 (68,57%)	132 (31,43%)	420	0 (0,00%)	1 (100,00%)	1
Osetalmivir	262 (70,62%)	109 (29,38%)	371	26 (52,00%)	24 (48,00%)	50
Glicocorticóide	140 (68,63%)	64 (31,37%)	204	148 (68,20%)	69 (31,80%)	217
Heparina	94 (60,65%)	61 (39,35%)	155	194 (72,93%)	72 (27,07%)	266
Terapia com Plasma	263 (67,96%)	124 (32,04%)	387	25 (73,53%)	9 (26,47%)	34
Antibioticoterapia	66 (64,71%)	36 (35,29%)	102	222 (69,59%)	97 (30,41%)	319
Outra proposta terapêutica	245 (71,22%)	99 (28,78%)	344	43 (55,84%)	34 (44,16%)	77

Tabela 14. Análise dos sinais de dispneia nas primeiras 24h de internação com relação ao desfecho.

Intensidade	Não grave (risco absoluto)	Grave (risco absoluto)
Ausência de dispneia	130 (72,22%)	50 (27,78%)
Dispneia leve 1+/4+	88 (89,80%)	10 (10,20%)
Dispneia 2+/4+	55 (64,71%)	30 (35,29%)
Dispneia 3+/4+	9 (32,14%)	19 (67,86%)
Dispneia 4+/4+	6 (20,00%)	24 (80,00%)

Tabela 15. Análise do estado geral à admissão hospitalar.

Estado Geral	Não grave (risco absoluto)	Grave (risco absoluto)
Bom	95 (77,24%)	28 (22,76%)
Levemente Comprometido	122 (85,31%)	21 (14,69%)
Moderadamente Comprometido	50 (57,47%)	37 (42,53%)
Gravemente Comprometido	8 (19,51%)	33 (80,49%)
Estado crítico/PCR abortada	0 (0,00%)	1 (100,00%)
Não preenchido	11 (42,31%)	15 (57,69%)

Com relação aos dados utilizados para treinar o classificador baseado em AM, a **Tabela 16** descreve quais variáveis foram utilizadas, bem como a distribuição entre os diferentes valores após categorização das variáveis numéricas. A **Tabela 17** descreve os dados após imputação pela média. Cabe reforçar que com a categorização das variáveis numéricas, de acordo com o descrito na sessão 3.2, seus valores foram trocados por valores inteiros que representam o quanto esta variável se desvia dos valores de referência ou da média dos pacientes no estudo. Assim, ao invés de utilizar o valor real da

hemoglobina, foram utilizadas categorias numeradas de 0 a 6, indo de um valor extremamente baixo até um valor elevado. Outras variáveis possuem menos classes, como o número de linfócitos (0 para reduzido, 1 para normal e 2 para elevado) e a presença de Diabetes Mellitus prévio (0 para não e 1 para sim). Após a imputação dos dados (**Tabela 17**), observa-se ausência de valores não preenchidos (pois estes foram substituídos pela média) e alteração na distribuição entre as categorias. Foram os valores de cada categoria que foram passados ao classificador, um novo paciente necessitaria ter seus exames categorizados utilizando o mesmo método antes de ser classificado.

4.4 RESULTADOS DOS CLASSIFICADORES

Durante 100 simulações, os 7 conjuntos de dados construídos neste estudo foram utilizados para o treinamento de 27 algoritmos de ML utilizando a biblioteca *Lazy Predict*. Em cada simulação os grupos de treinamento e teste foram separados de forma aleatória, na proporção de 80% e 20%, respectivamente, e a acurácia de cada simulação foi computada. Foi calculada então a acurácia média para cada classificador e para cada conjunto de dados, resultando nos valores mostrados na **Tabela 18**. Assim foi possível determinar que a maior acurácia média foi com o Classificador de Vetor de Suporte (SVC), para o conjunto de dados de 28 variáveis (conjunto F, sem imputação) e que os sete melhores classificadores para este conjunto foram *SVC*, *NuSVC*, *Random Forest Classifier*, *Ridge Classifier CV*, *Ridge Classifier*, *Extra Trees Classifier* e *Linear Discriminant Analysis*. O conjunto de dados com melhor resultado foi o conjunto F, apenas este conjunto foi utilizado nas etapas seguintes.

Tabela 18. Acurácia média por classificador (executado pela biblioteca *Lazy Predict*) e conjunto de dados

Classificador	Acurácia Média						
	23 variáveis	24 variáveis	25 variáveis	26 variáveis	27 variáveis	28 variáveis	28 variáveis com imputação
<i>XGBoost Classifier</i>	74,62%	77,25%	77,27%	77,32%	77,63%	77,20%	72,87%
<i>Decision Tree Classifier</i>	70,13%	70,13%	69,94%	70,19%	70,82%	71,44%	66,07%
<i>Logistic Regression</i>	74,83%	77,51%	78,06%	77,63%	77,65%	77,72%	75,02%
<i>AdaBoost Classifier</i>	73,77%	77,22%	75,87%	76,32%	76,30%	76,32%	72,72%
<i>Bagging Classifier</i>	75,14%	77,33%	78,21%	77,54%	77,84%	77,08%	74,55%
<i>Linear Discriminant Analysis</i>	74,70%	77,90%	78,76%	78,16%	78,40%	78,38%	75,46%
<i>Linear SVC</i>	74,80%	77,63%	78,32%	77,68%	77,77%	77,34%	75,09%
<i>SVC</i>	76,30%	79,12%	79,90%	79,44%	80,12%	80,78%	77,27%
<i>Passive Aggressive Classifier</i>	67,48%	67,96%	70,24%	70,93%	69,42%	69,94%	67,05%
<i>NuSVC</i>	75,07%	78,84%	79,52%	79,09%	79,75%	80,14%	76,35%
<i>Nearest Centroid</i>	68,04%	70,99%	71,95%	71,04%	70,58%	71,54%	69,78%
<i>LGBM Classifier</i>	74,49%	77,33%	77,23%	78,26%	78,40%	77,46%	73,36%
<i>Ridge Classifier</i>	74,68%	78,45%	78,84%	78,32%	78,51%	78,74%	75,52%
<i>Ridge Classifier CV</i>	74,75%	78,42%	78,84%	78,35%	78,77%	79,04%	75,75%
<i>Random Forest Classifier</i>	76,20%	78,97%	80,18%	79,30%	80,33%	80,10%	76,74%
<i>Gaussian NB</i>	70,03%	71,70%	72,94%	71,00%	71,56%	71,78%	69,52%
<i>Calibrated Classifier CV</i>	74,41%	77,88%	78,08%	76,67%	77,68%	76,90%	75,65%
<i>Bernoulli NB</i>	73,07%	74,36%	76,16%	75,68%	74,84%	76,10%	72,64%
<i>Extra Trees Classifier</i>	74,83%	77,72%	78,61%	78,49%	79,51%	79,56%	75,60%
<i>Quadratic Discriminant Analysis</i>	71,46%	72,51%	73,31%	72,04%	70,96%	68,82%	70,29%
<i>Extra Tree Classifier</i>	65,07%	69,19%	70,56%	68,60%	67,40%	67,74%	65,91%
<i>Label Spreading</i>	67,75%	70,66%	74,06%	73,26%	73,35%	74,24%	67,69%
<i>Label Propagation</i>	67,72%	70,70%	74,03%	73,26%	73,35%	74,24%	67,68%
<i>KNeighbors Classifier</i>	73,00%	76,34%	77,65%	75,82%	75,70%	75,80%	73,19%
<i>Perceptron</i>	68,89%	70,13%	71,35%	70,82%	69,88%	70,80%	67,05%
<i>SGD Classifier</i>	68,28%	71,54%	72,08%	72,18%	71,98%	71,38%	68,38%
<i>Dummy Classifier</i>	68,68%	70,76%	71,16%	69,82%	71,30%	69,56%	68,32%

Utilizando a biblioteca *Scikit-Learn*, estes sete classificadores foram utilizados para novas 5.000 simulações, novamente com redistribuições aleatórias entre os grupos de treinamento e teste (validação cruzada de Monte Carlo). Esta biblioteca reúne implementações de algoritmos baseadas em trabalhos anteriores (47). Nestas simulações, apenas o conjunto de dados (*dataset*) com 28 variáveis sem imputação foi utilizado (conjunto F). As métricas de avaliação foram computadas para cálculo da média e margem de erro, para um Intervalo de Confiança de 95% (IC 95). Cabe resumir o princípio de cada um dos sete classificadores:

- *SVC*: o Classificador por Vetor de Suporte busca definir um hiperplano de decisão para separar os dois grupos com a maior margem possível (47,49);
- *NuSVC*: semelhante ao *SVC*, mas adiciona um parâmetro ν (ou nu), que controla o número de vetores de suporte (47,49);
- *Random Forest Classifier*: durante o treinamento são estabelecidas árvores de decisão que “votam” para decidir a qual classe o indivíduo pertence (30,47);
- *Extra Trees Classifier*: semelhante ao *Random Forest*, mas a construção das árvores de decisão se dá de forma mais aleatória, o que pode aumentar a generalização do classificador (47,50);
- *Ridge Classifier*: utiliza a regressão *Ridge* para determinar a qual classe o indivíduo pertence (47,51);
- *Ridge Classifier CV*: incorpora Validação Cruzada ao classificador *Ridge* para definir o melhor valor de regularização (47,51);
- *Linear Discriminant Analysis*: busca uma combinação linear das variáveis que separe melhor os grupos de interesse (32,47).

Além destes algoritmos, foram aplicados a este conjunto de dados 11 configurações diferentes de redes neurais densas por 100 simulações, novamente com reamostragens aleatórias do grupo de teste e treinamento. Estas simulações permitiram identificar qual topologia ou configuração de rede neural obteve o melhor desempenho. Para entender o processo de análise da melhor topologia, pode-se descrever o erro do classificador como uma função dos hiperparâmetros escolhidos para a rede neural, como o número de camadas

e o número de unidades por camada. Objetiva-se minimizar o erro, mas, por limitações de poder computacional, este processo não pode ser exaustivo e testar todas as possíveis combinações. Então, foram definidas 11 combinações com pequenas variações e estas topologias foram submetidas a 100 simulações cada. A acurácia média para cada topologia foi registrada e aquela com melhor resultado foi selecionada.

A arquitetura de rede neural com melhor desempenho foi a com uma camada de 32 unidades, com ativação *ReLU* (Unidade Linear Retificada) e uma segunda camada com apenas uma unidade, novamente com ativação *ReLU*. Entre estas duas camadas, foi adicionado um *Dropout* de modo que, durante o treinamento, 20% das unidades da primeira camada são ignoradas. A saída da última camada determina a qual classe o indivíduo pertence. A lista abaixo descreve as configurações de redes neurais testadas, omitindo as camadas de entrada (os dados clínicos) e saída (última camada com uma unidade) por serem as mesmas em todos os cenários. As configurações estão no formato (número de unidades, taxa de *Dropout*), onde cada parênteses representa uma camada oculta.

1. (32 unidades, 20% *Dropout*);
2. (32 unidades, 20% *Dropout*), (16 unidades, 20% *Dropout*);
3. (32 unidades, 20% *Dropout*), (16 unidades, 20% *Dropout*), (8 unidades, 20% *Dropout*);
4. (16 unidades, 20% *Dropout*);
5. (8 unidades, 20% *Dropout*);
6. (8 unidades, 20% *Dropout*), (4 unidades, 20% *Dropout*);
7. (64 unidades, 20% *Dropout*), (32 unidades, 20% *Dropout*), (16 unidades, 10% *Dropout*), (8 unidades, sem *Dropout*);
8. (8 unidades, sem *Dropout*);
9. (16 unidades, sem *Dropout*), (8 unidades, sem *Dropout*);
10. (16 unidades, sem *Dropout*), (8 unidades, sem *Dropout*), (4 unidades, sem *Dropout*);
11. (8 unidades, sem *Dropout*), (4 unidades, sem *Dropout*).

A melhor topologia, primeira da lista acima, foi então submetida ao mesmo procedimento de teste que os demais classificadores, com 5.000 simulações.

A **Tabela 19** descreve os resultados de cada algoritmo após os 5.000 treinamentos com reamostragens aleatórias dos grupos de treinamento e teste. Verificou-se valores muito semelhantes entre os diferentes algoritmos e uma margem de erro muito pequena em torno da acurácia média (resultado das 5000 simulações). A alta especificidade, aliada a um valor preditivo positivo significativo da Classificação por Vetores de Suporte (SVC) indicam uma alta taxa de acertos quando o algoritmo classifica o paciente como grave. Já os algoritmos Classificador Ridge, Ridge CV e Discriminante Linear mantêm a acurácia enquanto apresentam melhores resultados ao identificar pacientes que não evoluirão para a forma grave. Já a rede neural apresentou baixa sensibilidade e menor acurácia que os demais métodos. Uma métrica importante para a avaliação de classificadores, a Área Sob a Curva ROC (AUC) indica que os classificadores testados tiveram desempenho menos satisfatório para a distinção entre as classes. Uma exceção foi o Random Forest, com AUC médio de 0,91 e acurácia média de 80%.

Tabela 19. Resultados de cada classificador após 5000 simulações, com Intervalo de Confiança de 95%. Legenda AUC: área sob a curva ROC.

	Acurácia	Escore F1	Valor Preditivo Positivo	Valor Preditivo Negativo	Sensibilidade	Especificidade	AUC
SVC	0,79 ± 0,0014	0,54 ± 0,0031	0,87 ± 0,0036	0,78 ± 0,0016	0,40 ± 0,0031	0,97 ± 0,0008	0,79 ± 0,0020
NuSVC	0,79 ± 0,0015	0,56 ± 0,0032	0,83 ± 0,0037	0,79 ± 0,0018	0,44 ± 0,0036	0,96 ± 0,0010	0,78 ± 0,0022
Random Forest Classifier	0,80 ± 0,0014	0,58 ± 0,0031	0,81 ± 0,0039	0,79 ± 0,0017	0,46 ± 0,0034	0,95 ± 0,0011	0,91 ± 0,0014
Extra Trees Classifier	0,79 ± 0,0014	0,57 ± 0,0031	0,77 ± 0,0039	0,79 ± 0,0017	0,47 ± 0,0035	0,94 ± 0,0012	0,78 ± 0,0019
Ridge Classifier CV	0,80 ± 0,0015	0,63 ± 0,0028	0,75 ± 0,0037	0,82 ± 0,0017	0,56 ± 0,0034	0,91 ± 0,0014	0,77 ± 0,0020
Ridge Classifier	0,78 ± 0,0015	0,62 ± 0,0028	0,69 ± 0,0038	0,82 ± 0,0017	0,57 ± 0,0033	0,88 ± 0,0016	0,77 ± 0,0020
Linear Discriminant Analysis	0,78 ± 0,0015	0,61 ± 0,0027	0,67 ± 0,0036	0,82 ± 0,0017	0,58 ± 0,0034	0,87 ± 0,0017	0,77 ± 0,0020
Rede Neural Densa	0,76 ± 0,0016	0,45 ± 0,0034	0,81 ± 0,0050	0,76 ± 0,0018	0,33 ± 0,0033	0,96 ± 0,0012	0,77 ± 0,0018

5. DISCUSSÃO

Foi realizado um longo processo de pré-processamento de dados, de modo a identificar o máximo de erros de preenchimento ou duplicatas. Além disso, o processo de seleção de variáveis envolveu não só a busca por melhor desempenho dos classificadores, mas também como uma forma de minimizar a perda de indivíduos após a limpeza dos dados. A construção de sete conjuntos de dados diferentes, abarcando quantidades diferentes de variáveis e, conseqüentemente, indivíduos, permitiu verificar o efeito gerado no resultado dos classificadores.

Durante o pré-processamento, foi realizado um teste sobre o efeito da imputação para preencher dados incompletos. O método realizado foi de preencher os dados faltantes com a média calculada para todos os indivíduos, ou seja, aqueles que não tiveram a variável, como por exemplo a hemoglobina, medida receberam o valor médio para a mesma variável de todos que tiveram o exame coletado. A imputação permite que alguns algoritmos de AM, que não aceitam informações incompletas na entrada, sejam aplicados para todos os indivíduos do estudo. No entanto, observou-se que a imputação pela média reduziu a acurácia do classificador. Entende-se que os dados clínicos não são variáveis independentes entre si e, deste modo, outras formas de imputação de dados poderiam ser aplicadas com melhores resultados. A partir de uma análise multivariada, os exames coletados para determinado indivíduo poderiam ser utilizados para estimar o valor daqueles que estão faltando. Um exemplo possível seria a utilização da hemoglobina, frequência respiratória e sinais clínicos de dispnéia para estimar a saturação periférica de oxigênio (SpO_2) naqueles pacientes que não puderam ter a oximetria aferida. Sugere-se a realização de novos estudos dedicados à identificação dos melhores métodos de imputação de dados clínicos para o desenvolvimento de classificadores.

Foi observado que a redução de variáveis de entrada, apesar de reduzir a perda de indivíduos (pela incompletude ou dados faltantes) e da complexidade dos sistemas classificadores, sacrifica a acurácia dos sistemas de classificação. Durante este trabalho foi necessário, através da experimentação, encontrar a melhor solução para este dilema. Os testes com os diferentes conjuntos de

dados auxiliaram a encontrar o melhor equilíbrio entre perdas de indivíduos e quantidade de dados clínicos utilizados, já que uma perda elevada de indivíduos sacrificaria o aprendizado dos classificadores, bem como a confiabilidade e capacidade de generalização dos resultados. A dependência de uma elevada quantidade de dados clínicos, aliada a incapacidade da maior parte dos algoritmos de lidar com incompletude, podem indicar desempenho inferior ao de um profissional treinado e reforçam sobre o uso da IA apenas como uma ferramenta e não uma forma de substituir pessoas.

Chama-se atenção à elevada quantidade de variáveis esparsas coletadas durante a coorte. Como observado na **Tabela 13**, diversas variáveis estiveram positivas em menos de 1% dos participantes do estudo, como a presença de arritmia ou doença valvar prévia. Esta característica cria um desafio particular ao se desenvolver um classificador. Isto porque, quanto mais variáveis são utilizadas na entrada, mais complexo será o classificador. Ainda que se imagine um simples questionário para diagnóstico de determinada doença, quanto mais perguntas forem necessárias para obter o diagnóstico, mais difícil será sua implementação. Classificadores complexos são mais propensos a sobreajuste (ou *overfitting*), o que reduz sua capacidade de generalizar para dados e novos pacientes (38,39). Ao mesmo tempo, alguns dados clínicos, apesar de raros, podem ser excelentes marcadores de desfecho desfavorável em certas patologias e sua eliminação da análise poderia comprometer a qualidade da análise final. Pensando nisso, o conhecimento clínico dos sinais e sintomas, enquanto potenciais preditores de gravidade (5–7) foi levado em consideração na seleção de variáveis, em conjunto com a união de diferentes variáveis com o mesmo significado clínico (a exemplo da variável que representava a presença de qualquer doença pulmonar prévia).

Com relação aos classificadores baseados em IA, deve-se levar em consideração que um grande desbalanço entre as classes prejudica o desenvolvimento de sistemas de classificação (18). Neste trabalho, aproximadamente 69% dos participantes não evoluíram com complicações graves e, assim, um classificador que julgasse todos os indivíduos como “não graves” teriam uma considerável taxa de acerto. No entanto, diante dos resultados encontrados, é possível afirmar que foi possível desenvolver modelos

com acurácia satisfatória, sobretudo ao classificar um indivíduo como pertencente ao grupo “grave”, o que pode ser evidenciado pelo alto valor preditivo positivo e a alta especificidade.

Especificidade e sensibilidade são fundamentais para determinar a qualidade de um teste de triagem e, quando elevados, indicam baixo índice de erros de classificação. Apesar da baixa sensibilidade dos classificadores desenvolvidos, a elevada especificidade indica uma pequena quantidade de falso positivos. Poucos falsos positivos são usualmente preferíveis, ainda que às custas da redução da sensibilidade, sobretudo quando um teste positivo poderia levar a intervenções desnecessárias (52,53).

Em um contexto de múltiplos pacientes internados e escassez de recursos, como durante a pandemia pela COVID-19, os valores preditivos atingidos seriam de extrema utilidade. O valor preditivo positivo de 87% atingido pelo SVC, aliado a uma pequena quantidade de falsos positivos, permite um melhor direcionamento de recursos e intervenções, como vigilância ou transferência para outra unidade de saúde (53). É importante ressaltar que apenas os dados da admissão foram realizados para esta análise e que o acompanhamento clínico ao longo da internação, bem como a realização de novos exames, poderia permitir uma reclassificação dos pacientes e aumentar o desempenho dos sistemas. Além disso, como os dados utilizados vieram de um estudo de coorte prospectivo, em novos estudos os classificadores e análises de dados podem ser desenvolvido em paralelo com a coleta dos dados, guiando a coorte na direção das variáveis mais importantes.

Ao analisar o custo-efetividade do uso dos classificadores desenvolvidos, deve-se levar em consideração que os exames laboratoriais e avaliação clínica realizados são indispensáveis para o adequado acompanhamento do paciente internado, de modo que o uso do classificador não exigiria gastos adicionais além de um sistema informático capaz de gerar os resultados. A utilização destes classificadores também não exige a realização de exames adicionais, diferentes daqueles fundamentais para o acompanhamento médico.

Nos últimos anos, diversos estudos têm sido publicados abordando a predição de evolução para desfechos desfavoráveis na COVID-19 a partir de

sistemas de IA (15,17,19,20,22,23). Um dos problemas encontrados com a literatura sobre AM e IA no diagnóstico de doenças ou avaliação de prognóstico foram as métricas de desempenho utilizadas. Poucos trabalhos publicaram os valores preditivos positivo e negativo, alguns não publicaram nem mesmo a sensibilidade e especificidade (20). Entende-se que um novo estudo clínico é necessário para determinação precisa destas métricas, utilizando dados de novos pacientes, inéditos aos sistemas desenvolvidos, mas considerou-se essencial o cálculo destas probabilidades para avaliar a viabilidade dos classificadores.

Foram publicados sistemas com acurácias variando de 74,4% a 95,20% e AUC entre 0,66 e 0,997 (19,20). No entanto, diferentemente deste trabalho, diversos estudos utilizaram dados coletados em pesquisas retrospectivas e unicêntricas, o que pode comprometer a capacidade de generalização dos classificadores para novos pacientes e reduz sua validade externa (20).

Kamel et al. (2023) investigaram a capacidade de diferentes algoritmos baseados em AM para prever desfecho desfavorável pela COVID-19, usando dados clínicos da admissão e exames laboratoriais (23). Ao comparar diferentes conjuntos de dados, determinaram que as variáveis hematológicas apresentaram a maior capacidade de predição de desfecho, mas a união de todos os dados coletados resultou em melhor acurácia. Diversos algoritmos testados apresentaram acurácia de 90%, com excelente sensibilidade, especificidade e AUC. Os dados foram coletados em uma coorte retrospectiva, que incluiu 485 pacientes, com imputação de dados incompletos pela média. No entanto, somente as variáveis que não foram coletadas em até 10 pacientes foram consideradas. Além disso, dados da internação hospitalar, e não somente da admissão, foram utilizados no classificador final, e melhor performance. Neste estudo a classe de pacientes graves representava apenas 20 a 23% do total (mortes e internação em UTI, respectivamente), com desbalanço ainda maior que na coorte utilizada neste estudo.

Ramírez-Del Real et al. (2022, v. 10) também atingiram resultados satisfatórios ao utilizar dados demográficos, clínicos e laboratoriais para prever mortalidade pela COVID-19 (22). Foram coletados dados de indivíduos saudáveis e estes foram acompanhados para avaliar o desfecho caso

contraíssem a COVID-19. Atingiram uma acurácia de 90,41%, com valores preditivos positivo e negativo de 94,28% e 87,36%, respectivamente. Este estudo sugere a possibilidade de identificar indivíduos mais vulneráveis a desfechos desfavoráveis, mesmo antes de contrair a doença.

Por utilizar dados coletados em um estudo multicêntrico, não seria impreciso supor que o sistema desenvolvido teria uma maior capacidade de generalização. Isto porque participaram do estudo unidades de saúde com diferentes capacidades, densidade tecnológica e acesso a recursos. Assim, o sistema poderia ser utilizado tanto em hospitais com menor estrutura quanto em unidades mais completas. Também cabe reforçar que, utilizados dados coletados em uma coorte prospectiva, o que garante maior confiabilidade sobre a qualidade dos dados, apesar de acrescentar limitações técnicas, como de coletar todas as variáveis em todos os participantes. Não foram encontrados outros estudos que utilizaram apenas informações clínicas e demográficas coletadas prospectivamente.

Outros trabalhos mostraram também resultados satisfatórios ao desenvolver preditores que adiciona exames de imagem, como Tomografia Computadorizada, aos dados clínicos para classificar gravidade pela COVID-19 (17). A coorte descrita neste trabalho coletou também radiografias de tórax no leito, que poderá ser aliada aos dados clínicos com objetivo de aprimorar os resultados em trabalho futuro. No entanto, uma vantagem da nossa pesquisa é que, ao utilizar poucas variáveis, resultantes de exames simples, permitiu-se o uso do sistema de predição mesmo em unidades de saúde com baixa densidade tecnológica. Isso porque estima-se que, em 2020, apenas cerca de 15% dos municípios brasileiros contavam com Tomografia Computadorizada (54).

Entende-se que a vacinação e o surgimento de novas variantes do SARS-CoV-2 alteram o comportamento da doença e o número de complicações graves. Foi observada a redução nos casos graves da COVID-19 em indivíduos vacinados e aumento da transmissão com o surgimento de novas variantes (6) e isto torna necessária a coleta de novos dados e atualização constante dos classificadores.

Este estudo validou a metodologia para o desenvolvimento de preditores de morbiletalidade, partindo da coleta de dados até o treinamento de algoritmos baseados em AM. Esta metodologia pode ser aplicada para outras doenças, sobretudo aquelas cujo desfecho depende da relação entre múltiplas variáveis.

6. CONCLUSÕES

Foi desenvolvido um algoritmo baseado em IA que obteve acurácia de 80% na predição de desfechos graves utilizando apenas dados da admissão hospitalar, com valores preditivos positivo e negativo também próximos a 80%. Frente aos resultados obtidos sugere-se que a metodologia proposta, que abrange, desde a coleta e processamento dos dados, até a aplicação de classificadores baseados em AM, demonstra ser uma abordagem passível de aplicação na COVID-19 e viável para lidar com novas doenças e crises de saúde. Esta abordagem tem o potencial de aprimorar significativamente a gestão de recursos e o atendimento aos pacientes.

Por se tratar de um trabalho multidisciplinar, é importante destacar a relevância de ir além da aplicação de algoritmos de pré-processamento e classificação, incorporando conhecimentos médicos. A união da visão médica e computacional permite uma análise mais adequada sobre viabilidade, aplicabilidade e validade do sistema de classificação.

Apesar de encontrar na literatura trabalhos com melhor desempenho na predição do desfecho, foram utilizados conjuntos de dados oriundos de coorte prospectiva e multicêntrica, realizada durante o auge das limitações de recursos públicos e privados. Estas limitações impactaram a qualidade dos dados, sobretudo em se tratando de informações faltantes ou incompletude, mas reforçam a qualidade do sistema classificador que foi desenvolvido, por se aproximar da realidade dos sistemas de saúde. Além disso, pela adequação às limitações reais, e por abranger indivíduos de diferentes situações socioeconômicas e perfis demográficos, o sistema desenvolvido tem maior usabilidade prática e pode abranger diferentes níveis de unidades de saúde.

Comprova-se, então, por meio dos resultados obtidos e revisão da literatura, que sistemas baseados em AM podem ser utilizados como preditores de gravidade em pacientes internados com COVID-19, ainda que utilizando apenas os dados da admissão hospitalar.

7. REFERÊNCIAS BIBLIOGRÁFICAS

1. Sott MK, Bender MS, da Silva Baum K. Covid-19 Outbreak in Brazil: Health, Social, Political, and Economic Implications. *International Journal of Health Services*. 1º de outubro de 2022;52(4):442–54.
2. Mohamed K, Rodríguez-Román E, Rahmani F, Zhang H, Ivanovska M, Makka SA, et al. Borderless collaboration is needed for COVID-19—A disease that knows no borders. *Infect Control Hosp Epidemiol*. 22 de outubro de 2020;41(10):1245–6.
3. World Health Organization. WHO Coronavirus (COVID-19) Dashboard [Internet]. 2023 [citado 7 de outubro de 2023]. Disponível em: <https://covid19.who.int/>
4. Luo X, Lv M, Zhang X, Estill J, Yang B, Lei R, et al. Clinical manifestations of COVID-19: An overview of 102 systematic reviews with evidence mapping. *J Evid Based Med*. 1º de setembro de 2022;15(3):201–15.
5. Ou M, Zhu J, Ji P, Li H, Zhong Z, Li B, et al. Risk Factors of Severe Cases with COVID-19: A Meta-Analysis. *Epidemiol Infect*. 2020;148(e175).
6. Martono, Fatmawati F, Mulyanti S. Risk Factors Associated with the Severity of COVID-19. Vol. 30, *Malaysian Journal of Medical Sciences*. Penerbit Universiti Sains Malaysia; 2023. p. 84–92.
7. Fang X, Li S, Yu H, Wang P, Zhang Y, Chen Z, et al. Epidemiological, comorbidity factors with severity and prognosis of COVID-19: a systematic review and meta-analysis. *Aging*. 13 de julho de 2020;12(13):12493–503.
8. Buonsenso D, De Rose C, Pierantoni L. Doctors' shortage in adults COVID-19 units: a call for pediatricians. *Eur J Pediatr*. 16 de julho de 2021;180(7):2315–8.
9. Yoshioka-Maeda K, Iwasaki-Motegi R, Honda C. Preventing the dysfunction of public health centres responding to COVID-19 by focusing on public health nurses in Japan. Vol. 76, *Journal of Advanced Nursing*. Blackwell Publishing Ltd; 2020. p. 2215–6.
10. Kanji S, Burry L, Williamson D, Pittman M, Dubinsky S, Patel D, et al. Therapeutic alternatives and strategies for drug conservation in the intensive care unit during times of drug shortage: a report of the Ontario COVID-19 ICU Drug Task Force. *Canadian Journal of Anesthesia*. 1º de outubro de 2020;67(10):1405–16.
11. Sandhu P, Shah AB, Ahmad FB, Kerr J, Demeke HB, Graeden E, et al. Emergency Department and Intensive Care Unit Overcrowding and Ventilator Shortages in US Hospitals During the COVID-19 Pandemic, 2020-2021. *Public Health Reports*. 1º de julho de 2022;137(4):796–802.

12. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. *J Med Syst*. 1º de agosto de 2020;44(8).
13. Arpacı I, Huang S, Al-Emran M, Al-Kabi MN, Peng M. Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms. *Multimed Tools Appl*. 1º de março de 2021;80(8):11943–57.
14. Mohamed K, Rzymiski P, Islam MS, Makuku R, Mushtaq A, Khan A, et al. COVID-19 vaccinations: The unknowns, challenges, and hopes. Vol. 94, *Journal of Medical Virology*. John Wiley and Sons Inc; 2022. p. 1336–49.
15. Kim DK. Prediction Models for COVID-19 Mortality Using Artificial Intelligence. Vol. 12, *Journal of Personalized Medicine*. MDPI; 2022.
16. Rahman MDA, Hossain MS, Alrajeh NA, Gupta BB. A multimodal, multimedia point-of-care deep learning framework for COVID-19 diagnosis. *ACM Transactions on Multimedia Computing, Communications and Applications*. 1º de abril de 2021;17(1s).
17. Yu L, Shi X, Liu X, Jin W, Jia X, Xi S, et al. Artificial Intelligence Systems for Diagnosis and Clinical Classification of COVID-19. *Front Microbiol*. 27 de setembro de 2021;12.
18. Isgut M, Gloster L, Choi K, Venugopalan J, Wang MD. Systematic Review of Advanced AI Methods for Improving Healthcare Data Quality in Post COVID-19 Era. Vol. 16, *IEEE Reviews in Biomedical Engineering*. Institute of Electrical and Electronics Engineers Inc.; 2023. p. 53–69.
19. Shakibfar S, Nyberg F, Li H, Zhao J, Nordeng HME, Sandve GKF, et al. Artificial intelligence-driven prediction of COVID-19-related hospitalization and death: a systematic review. Vol. 11, *Frontiers in Public Health*. Frontiers Media SA; 2023.
20. Wang L, Zhang Y, Wang D, Tong X, Liu T, Zhang S, et al. Artificial Intelligence for COVID-19: A Systematic Review. Vol. 8, *Frontiers in Medicine*. Frontiers Media S.A.; 2021.
21. Napolitano F, Xu X, Gao X. Impact of computational approaches in the fight against COVID-19: An AI guided review of 17 000 studies. Vol. 23, *Briefings in Bioinformatics*. Oxford University Press; 2022.
22. Ramírez-Del Real T, Martínez-García M, Márquez MF, López-Trejo L, Gutiérrez-Esparza G, Hernández-Lemus E. Individual Factors Associated With COVID-19 Infection: A Machine Learning Study. *Front Public Health*. 2022;10:912099.
23. Kamel FO, Magadmi R, Qutub S, Badawi M, Badawi M, Madani TA, et al. Machine Learning-Based Prediction of COVID-19 Prognosis Using Clinical and Hematologic Data. *Cureus*. 9 de dezembro de 2023;

24. Kühl N, Schemmer M, Goutier M, Satzger G. Artificial intelligence and machine learning. *Electronic Markets*. 1º de dezembro de 2022;32(4):2235–44.
25. Haglin JM, Jimenez G, Eltorai AEM. Artificial neural networks in medicine. Vol. 9, Health and Technology. Springer Verlag; 2019.
26. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*. 21 de dezembro de 2019;19(1).
27. Faouzi Johann and Colliot O. Classic Machine Learning Methods. Em: Colliot O, organizador. *Machine Learning for Brain Disorders*. New York, NY: Springer US; 2023. p. 25–75.
28. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. *IEEE Intelligent Systems and their Applications*. 1998;13(4).
29. Vapnik V, Golowich SE, Smola A. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. *NIPS'96: Proceedings of the 9th International Conference on Neural Information Processing Systems*. 1996;281–7.
30. Breiman L. Random Forests. *Mach Learn*. 2001;45:5–32.
31. Quinlan JR. Induction of Decision Trees. *Mach Learn*. 1986;1:81–106.
32. Tharwat A, Gaber T, Ibrahim A, Hassanien AE. Linear discriminant analysis: A detailed tutorial. *AI Communications*. 2017;30(2):169–90.
33. Grossi E, Buscema M. Introduction to artificial neural networks. Vol. 19, *European Journal of Gastroenterology and Hepatology*. 2007. p. 1046–54.
34. Rodvold DM, McLeod DG, Brandt JM, Snow PB, Murphy GP. Introduction to artificial neural networks for physicians: Taking the lid off the black box. *Prostate*. 2001;46(1):39–44.
35. Buscema M. A brief overview and introduction to artificial neural networks. Em: *Substance Use and Misuse*. 2002. p. 1093–148.
36. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016.
37. Rasamoelina AD, Adjailia F, Sinčák P. A Review of Activation Function for Artificial Neural Network. Em: *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. 2020. p. 281–6.
38. Pansambal BH, Nandgaokar AB. Integrating Dropout Regularization Technique at Different Layers to Improve the Performance of Neural Networks. Vol. 14, *IJACSA) International Journal of Advanced Computer Science and Applications*.
39. Salehin I, Kang DK. A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain. Vol. 12, *Electronics (Switzerland)*. Multidisciplinary Digital Publishing Institute (MDPI); 2023.

40. Zhou ZH. Machine Learning. Singapore: Springer Singapore; 2021.
41. Ying X. An Overview of Overfitting and its Solutions. J Phys Conf Ser. fevereiro de 2019;1168(2):22022.
42. Tipirneni S, Reddy CK. Self-Supervised Transformer for Sparse and Irregularly Sampled Multivariate Clinical Time-Series. ACM Trans Knowl Discov Data. 30 de julho de 2022;16(6).
43. Badireddy, M., & Baradhi, K. M. (2023). Chronic Anemia. In StatPearls. StatPearls Publishing.
44. Gloucestershire Hospitals NSH Founding Trust: Haematology reference ranges [Internet]. Disponível em: <https://www.gloshospitals.nhs.uk/our-services/services-we-offer/pathology/haematology/haematology-reference-ranges/> [Acesso em 12/05/2023].
45. NSH Blood Sciences Reference Range Database BSHC-REC-317b version 3.13 [Internet]. Disponível em https://www.leedsth.nhs.uk/wp-content/uploads/2023/09/Blood-Sciences-Reference-Range-Database_BSHC-REC-317-v3.pdf [Acesso em 12/05/2023].
46. Shan G. Monte Carlo cross-validation for a study with binary outcome and limited sample size. BMC Med Inform Decis Mak. 1º de dezembro de 2022;22(1).
47. Buitinck L, Louppe G, Blondel M, Fabien P, Mueller A, Olivier G, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–30.
48. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: A System for Large-Scale Machine Learning. 12th USENIX Symposium on Operating Systems Design and Implementation. 2016;
49. Chang CC, Lin CJ. LIBSVM: A Library for support vector machines. ACM Trans Intell Syst Technol. abril de 2011;2(3).
50. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. abril de 2006;63(1):3–42.
51. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970;12(1).
52. Herman C. What Makes a Screening Exam “Good”? Virtual Mentor [Internet]. 2006;8(1):34–37.
53. Trevethan R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. Front Public Health. 20 de novembro de 2017;5.
54. Pereira E. TOMÓGRAFOS X COVID-19 Análise da distribuição espacial de Tomógrafos no Brasil em tempos de pandemia. 2020. Projeto de extensão do Centro de Ensino Unificado do Distrito Federal (UDF).