UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MESTRANDO

WALLACE DUARTE DE HOLANDA

ORIENTADOR

Prof. Dr. LENARDO CHAVES E SILVA

Modelos de Aprendizado de Máquina para a Predição do Agravamento do Quadro Clínico de Pacientes com a COVID-19

Mossoró-RN

2022

WALLACE DUARTE DE HOLANDA

Modelos de Aprendizado de Máquina para a Predição do Agravamento do Quadro Clínico de Pacientes com a COVID-19

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação - associação ampla entre a Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semi-Árido, para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Lenardo Chaves e Silva

Mossoró-RN

2022

Catalogação da Publicação na Fonte.
Universidade do Estado do Rio Grande do Norte.

H722m Holanda, Wallace Duarte de
Modelos de Aprendizado de Máquina para a Predição do Agravamento do Quadro Clínico de Pacientes com a COVID-19. / Wallace Duarte de Holanda. - Mossoró, 2022.
96p.

Orientador(a): Prof. Dr. Lenardo Chaves Silva.
Dissertação (Mestrado em Programa de Pós-Graduação em Ciência da Computação). Universidade do Estado do Rio Grande do Norte.

1. Aprendizado de Máquina. 2. COVID-19. 3. Risco de Mortalidade. 4. Risco de Internação. 5. Modelo Preditivo. I. Silva, Lenardo Chaves. II. Universidade do Estado do Rio Grande do Norte. III. Título.

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC´s) foi desenvolvido pela Diretoria de Informatização (DINF), sob orientação dos bibliotecários do SIB-UERN, para ser adaptado às necessidades da comunidade acadêmica UERN.

UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE

UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

WALLACE DUARTE DE HOLANDA

MODELOS DE APRENDIZADO DE MÁQUINA PARA A PREDIÇÃO DO AGRAVAMENTO DO QUADRO CLÍNICO DE PACIENTES COM A COVID-19

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.

APROVADA EM: 02/12/2022

Prof. Dr. Lenardo Chaves e Silva
Orientador e Presidente

Prof. Dr. Álvaro Alvares de Carvalho César Sobrinho
Examinador Externo - Universidade Federal do Agreste de Pernambuco - UFAPE

Prof. Dr. Leiva Casemiro Oliveira
Examinador Interno - Universidade Federal Rural do Semi-Árido – UFERSA

Prof. Dr. Sebastião Emídio Alves Filho
Examinador Interno - Universidade do Estado do Rio Grande do Norte – UERN

Aos meus pais, Antônio e Edilene.

Agradecimentos

Antes de qualquer coisa, agradeço ao meu bom Deus, que ao longo desta jornada sempre me deu forças para persistir em busca dos meus sonhos. Obrigado, Senhor, por ouvir minhas orações e me fazer acreditar que vossos planos são perfeitos.

Agradeço aos meus pais, Antonio e Edilene, por tornarem possível a realização deste sonho. Obrigado por sempre confiarem em mim, por me darem forças e por serem a minha fortaleza.

Agradeço ao meu professor Lenardo Chaves, por todo suporte e auxílio do início ao final do mestrado. Obrigado por toda paciência, respeito, cooperação e por sempre acreditar no meu potencial. Sou e serei eternamente grato!

Agradeço aos professores Álvaro Alvares, Leiva Casemiro e Sebastião Alves, por se disponibilizarem a compor a Banca Examinadora e destacarem os pontos de melhorias deste trabalho.

Agradeço também a minha namorada e companheira de jornada, Adriana Ferreira, por todo amor, carinho e ajuda em todos os momentos, amo-te!

Agradeço a todos os meus professores e colegas do mestrado por todos os períodos de discussão de ideias e contribuições para o trabalho.

Mas os que confiam no SENHOR renovam suas forças;
voam alto, como águias. Correm e não se cansam,
caminham e não desfalecem.
Isaías 40:31.

Resumo

Desde o início da pandemia do novo coronavírus, houve um aumento nos índices de internação e mortalidade em todo o mundo. Os altos índices podem ser compreendidos diante da disseminação da doença e a dificuldade em identificar os pacientes com maior risco de agravamento do quadro clínico. A partir de um Mapeamento Sistemático da Literatura (MSL) foi observado que em meio a essa dificuldade, a utilização de modelos preditivos baseados em aprendizado de máquina se caracteriza como uma forma de auxiliar no prognóstico e na alocação antecipada de recursos para o tratamento dos pacientes. Neste estudo foram desenvolvidos dois modelos, direcionados na predição da internação e mortalidade de pacientes com a COVID-19. Para isso, foram avaliados 14 algoritmos, com destaque aos algoritmos *AdaBoost*, *Logistic Regression*, *Random Forest* e *Gradient Boosting*, que alcançaram os melhores resultados. Para o treinamento dos modelos foram utilizados dados demográficos, histórico de vacinação, sintomas e comorbidades de pacientes com casos suspeitos de COVID-19 atendidos em hospitais de São Paulo. Ao avaliar o desempenho desses algoritmos, foi observado que os modelos desenvolvidos a partir do algoritmo *Gradient Boosting* obtiveram os melhores resultados, alcançando uma acurácia de 83% e AUC (*Area Under Curve*) de 0.89 na predição da mortalidade, e acurácia de 71% e AUC de 0.75 na predição da internação. Também foi identificado que a idade avançada, a falta de ar e ausência de vacinação foram os principais fatores atrelados ao agravamento do quadro clínico. Por fim, visando demonstrar como os modelos propostos poderiam ser utilizados na prática pelos profissionais da Saúde, uma aplicação *Web* foi desenvolvida.

Palavras-chave: Aprendizado de Máquina, COVID-19, Risco de Mortalidade, Risco de Internação, Modelo Preditivo.

Abstract

Since the beginning of the new coronavirus pandemic, there has been an increase in hospitalization and mortality rates worldwide. The high rates can be understood through the spread of COVID-19 and the difficulty in identifying patients with a higher risk of worsening clinical status. From a Systematic Literature Mapping (SLM), it was observed that using predictive models based on machine learning algorithms is a way to assist in the prognosis and early allocation of resources for the treatment of potential patients. In this study, two predictive models were developed to predict hospitalization and mortality in patients with COVID-19. To this end, 14 algorithms were evaluated, with emphasis on the algorithms *AdaBoost*, *Logistic Regression*, *Random Forest* and *Gradient Boosting*, which achieved the best results. To train the models, demographic data, vaccination history, symptoms, and comorbidities of patients with suspected cases of COVID-19 from hospitals in São Paulo were used. When evaluating the performance of the algorithms, it was noted that the models developed from the Gradient Boosting algorithm obtained the best results, reaching an accuracy of 83% and an AUC (Area Under Curve) of 0.89 in predicting mortality, and accuracy of 71% and an AUC of 0.75 in predicting hospitalization. Moreover, it was also identified that advanced age, breathlessness, and lack of vaccination were the main factors associated with worsening the clinical condition. Finally, a Web application was developed to demonstrate how healthcare professionals could use the proposed models in practice.

Keywords: Machine Learning, COVID-19, Mortality Risk, Hospitalization Risk, Predictive Model.

Lista de ilustrações

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

ANN	*Artificial Neural Network*
AUC	*Area Under the Curve*
CE	Critério de Exclusão
CI	Critério de Inclusão
DNN	*Deep Neural Network*
DT	*Decision Tree*
FN	Falso Negativo
FP	Falso Positivo
GB	*Gradient Boosting*
LE	Lista de Trabalhos Excluídos
LI	Lista de Trabalhos Incluídos
LR	*Logistic Regression*
MAE	*Mean Absolute Error*
MLP	*Multilayer Perceptron*
MSL	Mapeamento Sistemático da Literatura
QC	Questão Conceitual
QGP	Questão Geral de Pesquisa
OMS	Organização Mundial da Saúde
QP	Questão de Pesquisa
QPT	Questão Prática
QS	Questão Secundária
QT	Questão Tecnológica
RF	*Random Forest*

RMSE	*Root Mean Squared Error*
RT-Ag	*Rapid Test - Antigen*
RT-LAMP	*Reverse Transcription - Loop-mediated isothermal amplification*
RT-PCR	*Reverse Transcription - Polymerase Chain Reaction*
SHAP	*Shapley Additive exPlanations*
SMOTE	*Synthetic Minority Over-sampling Technique*
SP	São Paulo
SVM	*Support Vector Machine*
UTI	Unidade de Terapia Intensiva
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

Sumário

1 Introdução

A COVID-19 é a doença causada pelo novo coronavírus, denominado SARS-CoV-2 (WHO, 2020). O primeiro caso da doença foi registrado na China, em dezembro de 2019, e desde então, devido o alto índice de transmissão, casos de indivíduos com a COVID-19 foram registrados em todos os países do mundo, culminando em uma pandemia (ZHU *et al.*, 2020).

Os elevados índices de transmissão e mortalidade podem ser compreendidos por meio do aspecto clínico instável da doença quanto ao surgimento dos sintomas (SANTANA *et al.*, 2021). Devido essa característica, enquanto alguns pacientes infectados não manifestam sintomas, uma outra parcela sofre significativamente pelo agravamento do quadro clínico, culminando, em muitos casos, na morte desses indivíduos (GUAN *et al.*, 2020). Dessa forma, mesmo pacientes com quadro estável da doença podem vir a desenvolver, em um curto período de tempo, sintomas mais graves, dificultando o tratamento e o controle da COVID-19 (ISER *et al.*, 2020; SHEN *et al.*, 2022).

Diante disso, torna-se necessário identificar de maneira antecipada os pacientes diagnosticados com COVID-19 que apresentam a possibilidade do agravamento de seu quadro clínico (CEN *et al.*, 2020). A partir dessa constatação, seria possível auxiliar os gestores de saúde na alocação antecipada de recursos para o tratamento desses potenciais pacientes, tais como: profissionais, equipamentos e leitos de UTI (ASSAF *et al.*, 2020).

No âmbito da saúde, o diagnóstico antecipado é essencial para a administração dos pacientes, de modo a priorizar o atendimento e identificar a necessidade de medicação, equipamentos médicos e procedimentos clínicos, visando a aumentar as chances de sobrevivência (RAMALHO *et al.*, 2020; SILVA *et al.*, 2021; OLIVEIRA *et al.*, 2022).

Em meio a esta necessidade, a utilização de técnicas baseadas no Aprendizado de Máquina (AM) surge como solução (HOLANDA; SILVA; SOBRINHO, 2021; KHAN *et al.*, 2021). O AM é baseado no uso de algoritmos que possibilitam a predição de novas informações a partir de dados já existentes (MITCHELL, 1997). Na medicina, o AM tem sido utilizado corriqueiramente na construção de modelos preditivos, de modo a: *i)* fornecer um prognóstico de pacientes; *ii)* prever o avanço de doenças, e; a *iii)* detectar fatores de risco (SIDEY-GIBBONS; SIDEY-GIBBONS, 2019; LEE; CHOI; SHIN, 2021).

No entanto, a maioria dos modelos de AM desenvolvidos com foco na predição do agravamento do quadro clínico de pacientes com COVID-19: *i)* necessitam de dados de exames clínicos ou imagens de raio-x, fato esse que reduz a utilização prática em alguns casos devido a necessidade de realização dos exames; *ii)* avaliam o agravamento do quadro clínico em relação às chances de mortalidade ou internação, mas não em ambos os casos;

iii) não analisam o número de doses de vacinas tomadas pelos pacientes como um fator de agravamento do quadro clínico, e; *iv)* são pouco disponibilizados por meio de aplicativos ou páginas *Web*, o que impossibilita o uso dos modelos (HOLANDA; SILVA; SOBRINHO, 2021).

Diante disso, o objetivo deste estudo é propor e disponibilizar modelos de aprendizado de máquina com foco na predição do agravamento (internação ou óbito) do quadro clínico de pacientes com a COVID-19, a partir de dados demográficos, sintomas, presença de comorbidades e número de vacinas tomadas.

1.1 Questões de Pesquisa

O presente trabalho segue a abordagem *Design Science*, proposta por Wieringa (2014). Dessa forma, um conjunto de questões são definidas a fim de serem respondidas no decorrer da pesquisa. Diante desta definição, a Questão Geral de Pesquisa (QGP) a ser respondida é definida a seguir:

> **QGP: Como identificar a probabilidade de agravamento do quadro clínico de pacientes diagnosticados com a COVID-19 a partir de modelos preditivos?**

Com intuito de especificar a QGP, Wieringa (2014) sugere dividi-la em: Questões Conceituais (QC), Questões Tecnológicas (QT) e Questões Práticas (QPT). Com isso, a QGP é definida em questões mais aprofundadas, como descrito a seguir.

> **QC - Como construir um modelo de predição baseado em técnicas de aprendizado de máquina?**
>
> **QC1** O que é um modelo de predição?
>
> **QC2** Quais as abordagens encontradas na literatura para o desenvolvimento de modelos de predição?
>
> **QC3** Como essas abordagens são aplicadas tendo em vista os diferentes tipos de problemas?

> **QT - Quais os procedimentos necessários para o desenvolvimento de modelos preditivos?**
>
> **QT1** Quais linguagens de programação e/ou softwares podem ser usados no desenvolvimento do modelo preditivo?
>
> **QT2** Quais dados advindos dos pacientes diagnosticados com COVID-19 devem ser utilizados?

QT3 Quais algoritmos de aprendizado de máquina podem ser utilizados?

QPT - Como avaliar a qualidade do modelo preditivo?

QPT1 Quais métricas podem ser consideradas adequadas para a verificação dos resultados?

QPT2 Quais dados sobre os pacientes diagnosticados com COVID-19 são mais relevantes?

QPT3 Quais as alternativas existentes para possibilitar a disponibilização do modelo de predição?

1.2 Objetivos

O objetivo geral se resume em **propor modelos de aprendizado de máquina para a predição do agravamento do quadro clínico de pacientes com a COVID-19, considerando a probabilidade de internação ou óbito.** Para alcançar o objetivo geral, um conjunto de objetivos específicos devem ser contemplados, incluindo:

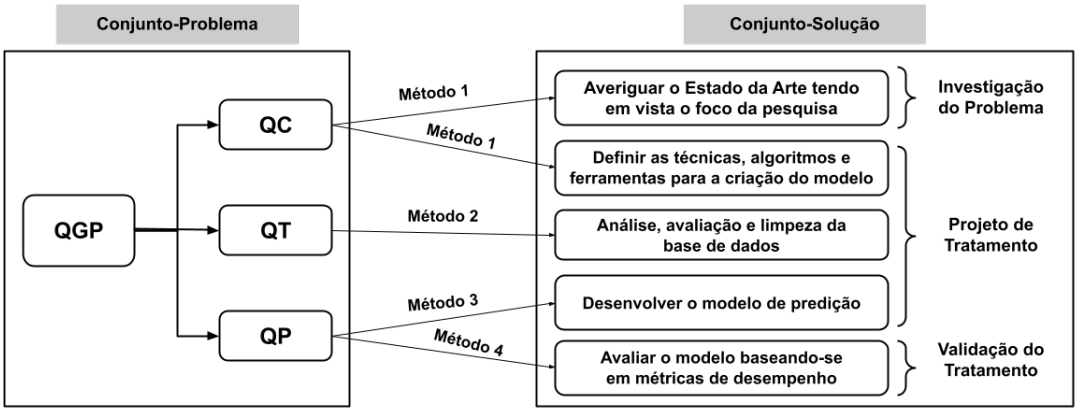
- investigar a literatura acerca do estado da arte para identificar trabalhos relacionados ao desenvolvimento de modelos de predição no âmbito da COVID-19, visando compreender as metodologias, técnicas, ferramentas e resultados alcançados;
- identificar quais desses trabalhos propõem modelos de predição do agravamento do quadro clínico de pacientes diagnosticados com a COVID-19;
- propor uma solução baseada em modelos de predição que possibilite a identificação da probabilidade de agravamento do quadro clínico de pacientes com COVID-19 a partir de informações sobre o estado de saúde dos mesmos;
- avaliar a eficiência dos modelos de predição desenvolvidos tendo em vista as métricas de desempenho existentes; e,
- disponibilizar o modelo desenvolvido de modo a auxiliar os profissionais de saúde na identificação dos pacientes com maior probabilidade de agravamento do quadro clínico.

1.3 Metodologia

A metodologia foi constituída a partir do *Design Science*, sendo dividida em duas partes: conjunto-problema e conjunto-solução. O conjunto-problema deste trabalho foi caracterizado através de uma Questão Geral de Pesquisa, sendo decomposta em: Questões

Conceituais, Questões Tecnológicas e Questões Práticas, conforme detalhado na Seção 1.1. Já em relação ao conjunto-solução, são especificados o conjunto de soluções propostas para alcançar as respostas necessárias às questões do conjunto-problema. Para isso, é necessário realizar as atividades de: i) investigação do problema, ii) projeto de tratamento e iii) validação do tratamento. Na Figura 1 é apresentado a concepção do processo.

Figura 1 – Metodologia de pesquisa na perspectiva do *Design Science*.



Fonte: Adaptada de Wieringa (2014).

Como apresentado na Figura 1, a metodologia de pesquisa adotada é caracterizada por um conjunto de métodos de pesquisa que associam o conjunto-problema ao conjunto-solução. No presente trabalho foram identificados quatro métodos, descritos a seguir:

Método 1: Mapeamento Sistemática da Literatura. Segundo Kai *et al.* (2008), um Mapeamento Sistemático Literatura (MSL) se caracteriza como um estudo que visa encontrar, avaliar e agregar as contribuições de estudos publicados em relação a uma questão de pesquisa específica. A partir dos trabalhos selecionados é possível avaliar os resultados alcançados e fornecer um panorama sobre a área de pesquisa.

Método 2: Pré-processamento dos Dados. Segundo Han, Kamber e Pei (2011), o pré-processamento dos dados se caracteriza como uma técnica direcionada no tratamento e remoção de aspectos que possam impactar na qualidade dos dados. Na maioria dos casos, as bases de dados são suscetíveis a dados ausentes e inconsistentes, tendo em vista o seu tamanho, organização e fonte de origem. Com isso, torna-se necessário a aplicação de técnicas de pré-processamento, a citar: limpeza, integração, redução e transformação dos dados.

Método 3: Aprendizado de Máquina. De acordo com Mitchell (1997), o Aprendizado de Máquina se resume no desenvolvimento de técnicas computacionais, bem como na construção de modelos preditivos. Ainda segundo Mitchell (1997), os modelos preditivos são desenvolvidos por meio de algoritmos de aprendizado de

máquina e se fundamentam em dados previamente analisados para fornecer indícios sobre a chance da ocorrência de um determinado evento.

Método 4: Avaliação de Desempenho. De acordo com VanderPlas (2016), a Avaliação de Desempenho de modelos preditivos é uma etapa fundamental na validação do seu funcionamento. Para a condução deste processo são utilizadas métricas de desempenho. As métricas são medidas quantificáveis usadas para analisar o o desempenho do modelo em diferentes perspectivas.

1.4 Motivação

No dia 11 de Março de 2020, a Organização Mundial da Saúde (OMS) declarou oficialmente a pandemia do novo coronavírus. Desde então, a COVID-19 vem impactando seriamente no sistema mundial de saúde, devido o elevado número de pessoas infectadas e a crescente demanda por profissionais e equipamentos médicos para tratar esses pacientes, tais como a alocação de leitos de Unidade de Terapia Intensiva (UTI) (NORONHA *et al.*, 2020). Atualmente já existe uma série de testes baseados em métodos diferentes que possibilitam identificar se um determinado paciente se encontra ou não com a doença, a citar:

- O RT-PCR (do inglês, *Reverse Transcription - Polymerase Chain Reaction*) é o mais comum e também o mais eficaz. Ele detecta o material genético do vírus em amostras respiratórias e, em razão da confiabilidade, é o mais indicado dentre os tipos de testes. Outra vantagem deste teste é o fato de ser possível detectar se o paciente está com a doença no momento da realização. No entanto, este teste costuma ser mais demorado que os outros tipos, tendo o resultado disponibilizado entre 48 e 72 horas (TEYMOURI *et al.*, 2021).
- O RT-LAMP (do inglês, *Reverse Transcription - Loop-mediated isothermal amplification*) é realizado de forma similar ao RT-PCR, pois detecta o DNA do microrganismo por meio de secreção das vias aéreas superiores. Além disso, os resultados do teste costumam sair um pouco mais rápidos que o RT-PCR, o que pode torna-lo mais indicado em alguns casos. No entanto, se comparado com o RT-PCR, o RT-LAMP possui uma taxa de eficácia ligeiramente menor (AMARAL *et al.*, 2021).
- O RT-Ag (do inglês, *Rapid Test - Antigen*), também conhecido como teste de antígeno, é um teste rápido que ao invés de detectar o material genético do vírus, como no RT-PCR, identifica as proteínas do vírus. O resultado deste teste é obtido em poucos minutos, o que torna extremamente útil em uma primeira avaliação. No entanto, os resultados não são tão precisos se comparado com os demais testes, o que pode gerar falsos negativos (SHUKLA *et al.*, 2022).

No entanto, mesmo com diversos avanços no contexto de identificação dos pacientes cometidos pela COVID-19, os profissionais de saúde ainda enfrentam dificuldades em identificar de forma antecipada os pacientes que possam desenvolver os sintomas mais graves da doença. De acordo com HEALTH (2022), a principal dificuldade está relacionada ao fato de que o vírus se comporta de forma instável em diferentes indivíduos.

Devido a essa característica, um paciente pode manifestar sintomas comuns da doença, tais como: febre, tosse seca, fadiga, perda de apetite, perda de olfato e dor no corpo. Mas após alguns dias, o mesmo paciente pode ter o seu estado clínico agravado a partir do desenvolvimento de sintomas mais intensos, a citar: febre alta, dor no peito, tosse excessiva e falta de ar (POUDEL *et al.*, 2021; HEALTH, 2022).

Razu *et al.* (2021) também destacam que além das dificuldades enfrentadas devido a instabilidade da doença em relação a manifestação dos sintomas, os profissionais de saúde possuem dificuldades na priorização e gerenciamento dos pacientes atendidos, devido: *i)* a proporção de profissionais de saúde em relação ao número de pacientes; *ii)* a elevada carga de trabalho; *iii)* escassez de equipamentos de proteção individual; *iv)* falta de incentivos, e; v) ausência de coordenação diante da quantidade de internados.

Nesse sentido, este trabalho pode auxiliar os profissionais de saúde na identificação prévia dos pacientes com COVID-19 que apresentem uma maior probabilidade de agravamento. Dessa forma, será possível contribuir na tomada de decisões antecipadas em relação ao uso de recursos médicos por parte dos gestores da área da saúde, com intuito de reduzir os impactos no sistema de saúde e minimizar as chances de mortalidade de pacientes.

1.5 Produção Científica

Evidenciando a relevância desta pesquisa, ao decorrer do seu desenvolvimento, dois trabalhos foram produzidos:

- Holanda, W. D.; Silva, L. C.; Sobrinho, A. A. C. C. *Estratégias Preditivas na Detecção do Agravamento do Quadro Clínico de Pacientes com COVID-19: Uma Revisão de Escopo*. In: Journal of Health Informatics (JHI), v. 13, n. 4, p. 128-132, 2021.
- Holanda, W. D.; Silva, L. C.; Sobrinho, A. A. C. C. *Estratégias Preditivas na Detecção do Agravamento do Quadro Clínico de Pacientes com COVID-19: Uma Revisão de Escopo*. In: XXII Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS), Teresina, 2022.

1.6 Organização do Documento

Além da Introdução, este documento está dividido em outros seis capítulos. No Capítulo 2 é apresentada a fundamentação teórica, destacando os principais conceitos abordados neste trabalho. No Capítulo 3 é descrito o protocolo do Mapeamento Sistemática da Literatura e seus resultados alcançados. No Capítulo 4 são detalhadas as etapas necessárias para o análise dos dados e desenvolvimento dos modelos de predição. No Capítulo 5 são apresentado e discutidos os resultados alcançados pelos algoritmos de aprendizado de máquina. No Capítulo 6 são apresentadas as conclusões alcançadas com esta pesquisa, limitações e qual a perspectiva de trabalhos futuros.

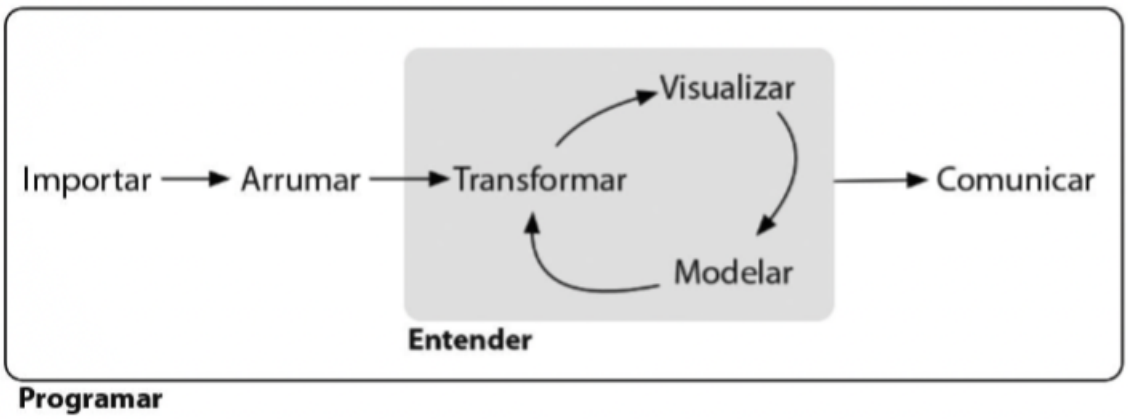
2 Fundamentação Teórica

Neste capítulo é apresentada a fundamentação teórica sobre as temáticas inerentes ao presente trabalho. Inicialmente, é apresentado o conjunto de etapas envolvidas no fluxo de Ciência de Dados (Seção 2.1). Em seguida, são ilustrados os principais conceitos que envolvem o Aprendizado de Máquina (Seção 2.2) e a Aprendizagem Supervisionada (Seção 2.3). Ao final, na Seção 2.4 são apresentadas as considerações finais do capítulo.

2.1 Ciência de Dados

Segundo Wickham e Grolemund (2017), a Ciência de Dados é caracterizada como a análise e exploração de dados, com intuito de compreender, extrair informações e gerar resultados a partir de dados brutos. Para isso, a Ciência de Dados utiliza a combinação de ferramentas, técnicas e algoritmos na descoberta de padrões e relacionamento entre as variáveis.

Figura 2 – Fluxo de etapas da Ciência de Dados.



Fonte: Wickham e Grolemund (2017).

Como apresentado na Figura 2, a importação é a etapa em que os dados são obtidos. Atualmente, as formas mais comuns de disponibilidade de dados são por meio de arquivos estruturados (.csv ou .xlsx) ou banco de dados. Após a importação, os dados são arrumados. Nesse momento será possível entender quais atributos estão disponíveis e em que formato foram armazenadas.

Em seguida, na etapa de transformação, a base de dados será avaliada com intuito de extrair novas informações (por exemplo: média dos valores e os valores que mais se repetem) e remover inconsistências (tais como: dados ausentes, formatos inadequados e

valores duplicados) que possam impactar nas etapas seguintes. A etapa de visualização se resume na apresentação gráfica dos dados. Neste momento será possível compreender de maneira visual a disposição, a distribuição e o estado atual dos dados.

Posteriormente, será possível utilizar o aprendizado de máquina na construção de um modelo, com intuito de identificar padrões e responder questões sobre os dados. Ao final, a etapa de comunicação é direcionada na exposição dos resultados para a comunidade científica e as conclusões obtidas ao longo de todo o processo de análise de dados.

Dessa forma, a Ciência de Dados se refere a um conjunto de etapas que envolvem a melhoria, análise e visualização dos dados, com objetivo de responder questões e gerar informações (HAN; KAMBER; PEI, 2011). Além disso, Cerri e Carvalho (2017) destacam que a Ciência de Dados possibilita a construção de modelos baseados em aprendizado de máquina que conseguem obter conhecimento a partir dos dados. Na seção a seguir serão apresentados os conceitos inerentes ao aprendizado de máquina.

2.2 Aprendizado de Máquina

Nos últimos anos, devido a crescente necessidade de processar volumes cada vez maiores de dados, tornou-se fundamental a busca por soluções computacionais eficientes e autônomas que fossem capazes de desenvolver hipóteses, funções ou resolver problemas a partir de experiências passadas (FACELI *et al.*, 2011). As soluções que seguem esse princípio são baseadas em aprendizado de máquina. Segundo Mitchell (1997), o aprendizado de máquina é uma área de pesquisa que visa o desenvolvimento de programas com a capacidade de aprender a executar uma determinada tarefa a partir de um conjunto de dados.

De acordo com Cerri e Carvalho (2017), um exemplo básico de aprendizado de máquina pode ser definido como um programa que possibilita a distinção de três variedades de flores. Dessa forma, um conjunto de dados contendo características botânicas pode ser utilizado na implementação de um modelo de aprendizado de máquina. Com isso, a partir de um processo de treinamento, o modelo aprenderá a identificar uma flor com base nas características contidas no conjunto de dados. Segundo Faceli *et al.* (2011) o aprendizado de máquina é categorizado em três tipos:

- **Aprendizado Supervisionado:** possibilita o desenvolvimento de modelos que conseguem aprender a partir de um conjunto de dados rotulados. Com isso, a partir deste tipo de aprendizagem será possível encontrar uma função que proporcione o mapeamento de novos exemplos, gerando assim, a predição de valores ainda não conhecidos;
- **Aprendizado Não-Supervisionado:** busca agrupar exemplos não rotulados através

da identificação de padrões ou tendências semelhantes ao conjunto de dados. Devido essa característica, este tipo de aprendizado é utilizado em tarefas descritivas, ou seja, ao receber um conjunto de dados de entrada sem a informação de saída, será possível encontrar grupos com propriedade similares, e;

- **Aprendizado Semi-Supervisionado:** possibilita o aprendizado a partir de dados rotulados e não rotulados. Este tipo de aprendizado é utilizado quando uma grande quantidade de dados está sendo manipulada, mas apenas alguns dados são rotulados. Dessa forma, é possível utilizar o aprendizado semi-supervisionado para aprender informações a partir dos dados não rotulados juntamente com os dados rotulados.

Neste trabalho será utilizado a abordagem de aprendizado supervisionado, que será detalhada na Seção 2.3.

2.3 Aprendizado Supervisionado

O aprendizado supervisionado se resume na predição de um valor de saída mediante um conjunto de dados de entrada (MITCHELL, 1997). Podem ser citados como exemplos: a predição do valor de um veículo dado o ano de seu lançamento, a classificação de frutas com base em suas características, e a predição da possibilidade de agravamento de doenças com base nos sintomas dos pacientes.

Além disso, o aprendizado supervisionado se caracteriza pela presença de um atributo especial que descreve o fenômeno de interesse, denominado de atributo alvo. A partir da definição deste atributo será possível treinar o algoritmo por meio de um conjunto de instâncias, com intuito de possibilitar a predição do atributo alvo de novas instâncias. As tarefas preditivas podem ser divididas em duas categorias: classificação e regressão (CERRI; CARVALHO, 2017).

A classificação é caracterizada pela atribuição de categorias predefinidas a dados. Por exemplo, um banco pode desenvolver um sistema para a classificação de seus clientes em duas categorias para fornecimento de empréstimo: SIM e NÃO. Para isso, baseando-se no histórico de crédito, emprego e salário de um determinado cliente (dados de entrada), o sistema aprenderá a distinguir os clientes para os quais o banco deve ou não deve fornecer um empréstimo (FACELI *et al.*, 2011).

A regressão se resume em prever o valor de uma variável numérica (atributo de saída), a partir de um conjunto de variáveis de entrada (atributos de entrada). Com isso, em uma perspectiva diferente da classificação, não será possível encontrar uma categoria associada, mas sim uma função que mapeie uma variável de entrada para um valor numérico de saída. Como exemplos de regressão, destacam-se a predição de valores de ações, predição de volume de chuvas e predição de preços de produtos (FACELI *et al.*, 2011).

Com isso, para que seja possível a criação de um modelo preditivo, seja baseado em classificação ou regressão, é necessário a realização dos processos de treinamento e teste. Para a realização dessas atividades são selecionados conjuntos de dados distintos. O conjunto de dados referente ao treinamento é utilizado no desenvolvimento do modelo, enquanto o conjunto de dados de teste é aplicado na avaliação do desempenho (REZENDE, 2003).

Bennett *et al.* (2022) destacam que os métodos mais utilizados na definição dos dados de treinamento e teste são o *hold-out* e *cross-validation*. O método *hold-out* consiste em dividir a base de dados em dois conjuntos, um para treinamento e outro para teste. Uma proporção muito comum deste método é separar 70% dos dados para treinamento e os 30% restante para teste. Em outra perspectiva, o método *cross-validation*, também chamado de *k-fold*, consiste em dividir a base de dados em k conjuntos de mesmo tamanho. Com essa definição, um conjunto de dados é utilizado para teste, enquanto os k - *1* são utilizados para o treinamento. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste. Na Seção a seguir são apresentadas algumas métricas utilizadas na avaliação do desempenho dos modelos de predição baseados em regressão ou classificação.

2.3.1 Métricas de Avaliação para Regressão

As métricas usadas para avaliar modelos baseados em regressão devem ser capazes de trabalhar em um conjunto de valores contínuos, tendo em vista a natureza dos dados. Dentre as métricas disponíveis, Wang e Lu (2018) destacaram a existência de duas: *Mean Absolute Error* e *Root Mean Squared Error*, vistas a seguir.

- ***Mean Absolute Error* (MAE)**: esta métrica possibilita a identificação do erro médio em um conjunto de previsões de tamanho N. Ou seja, o MAE é a diferença absoluta média entre os resultados observados (y_i) e os previstos ($\widehat{y}_i$) (Equação 2.1).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i) \tag{2.1}$$

- ***Root Mean Squared Error* (RMSE)**: esta métrica identifica o erro médio executado pelo modelo ao realizar a predição. Matematicamente, o RMSE é a raiz quadrada do erro quadrático médio de N observações, que é a diferença quadrática média entre os valores reais observados (y_i) e os valores previstos ($\widehat{y}_i$) pelo modelo (Equação 2.2). Dessa forma, quanto menor o RMSE, melhor é o modelo.

$$RMSE = \sqrt{\sum_{i=1}^{N}\frac{(y_i - \widehat{y}_i)^2}{N}} \tag{2.2}$$

Mesmo com algumas similaridades, as métricas apresentadas possibilitam a avaliação do modelo preditivo em diferentes perspectivas. Por exemplo, enquanto MAE pode ser apropriado para circunstâncias com poucos valores atípicos em relação aos demais, também chamados de *outliers*, o RMSE é constantemente utilizado em conjuntos de dados que contêm vários *outliers*.

2.3.2 Métricas de Avaliação para Classificação

O processo de avaliação de modelos baseados em classificação é definido pela quantidade de instâncias de teste classificadas. Nesse sentido, Burkov (2019) destaca que a matriz de confusão pode ser utilizada para visualizar o desempenho do modelo. A matriz de confusão é uma tabela que resume o sucesso do modelo de classificação em prever exemplos pertencentes a várias classes, sendo dividida em: Verdadeiros Positivos (VP), instâncias classificadas corretamente; Falsos Positivos (FP), instâncias classificadas como pertencentes a classe, mas na verdade não pertenciam; Falsos Negativos (FN), instâncias classificadas como não pertencentes a classe, mas na verdade pertenciam; e Verdadeiros Negativos (VN), instâncias classificadas como não pertencentes a classe, e que realmente não pertenciam. A matriz de confusão é ilustrada no Quadro 1.

Quadro 1 – Matriz de Confusão.

	Verdadeiro	**Falso**
Verdadeiro	VP	FP
Falso	FN	VN

Fonte: Burkov (2019).

Dessa forma, a matriz de confusão fornece indicadores da quantidade de ocorrências que um modelo preditivo teve para cada uma das quatro categorias. Além disso, as categorias apresentadas na matriz de confusão são utilizadas para calcular métricas de desempenho, como: Acurácia, Precisão, *Recall* e *F1-Score*.

- **Acurácia**: trata-se da proximidade de um resultado com o seu valor de referência. Dessa forma, quanto maior a acurácia, mais próximo da referência é o resultado encontrado (Equação 2.3).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.3}$$

- **Precisão**: é a métrica utilizada para avaliar a taxa de acerto na classificação do modelo preditivo. Ou seja, esta pontuação verifica a proporção de resultados que são Verdadeiro Positivo (VP). Para isso, a quantidade de valores Verdadeiro Positivo (VP) e Falso Positivo (FP) são analisados (Equação 2.4).

$$Precisão = \frac{VP}{(VP + FP)} \tag{2.4}$$

- ***Recall***: trata-se da pontuação atrelada a avaliação da quantidade de valores positivos esperados classificados corretamente. Para isso, a quantidade de valores Verdadeiro Positivo (VP) e Falso Negativo (FN) são usados para mensurar o valor do *Recall* (Equação 2.5).

$$Recall = \frac{VP}{(VP + FN)} \tag{2.5}$$

- ***F1-Score***: verifica a consistência entre as métricas de Precisão e *Recall* obtidas pelo classificador. Os valores desta pontuação estão dispostos na faixa de 0 e 1 (Equação 2.6).

$$F1 = 2 \times \frac{Precisão \times Recall}{Precisão + Recall} \tag{2.6}$$

- ***Area Under the Curve* (AUC)**: representa o grau ou medida de separabilidade. A partir deste valor é possível identificar o quanto o modelo é capaz de distinguir as classes. Dessa forma, quanto maior o valor de AUC, melhor o modelo é para prever as classes corretamente.

Tendo em vista a variedade de métricas existentes, é possível avaliar o modelo baseado em classificação em diferentes aspectos. Esse fato é importante pois um bom desempenho em uma métrica não garante a aceitação nas demais.

2.3.3 Algoritmos

Devido a relevância e a importância do aprendizado de máquina para o contexto de inúmeros problemas nas mais diferentes áreas, uma grande variedade de algoritmos pode ser encontrada. Nesta Seção são descritos alguns algoritmos de aprendizado supervisionado.

Decision Tree

Trata-se de um algoritmo de aprendizado supervisionado utilizado para problemas de classificação e regressão. Devido sua estrutura, esse algoritmo é geralmente representado visualmente por meio de uma árvore. O algoritmo se baseia na contínua subdivisão de um espaço amostral (raiz) em classes menores por meio de testes (nós). O espaço amostral é dividido até ser possível identificar um subconjunto homogêneo o suficiente para ser classificado como uma mesma classe, criando assim, um nó terminal (folha) (FACELI *et al.*, 2011). Para definir o melhor critério, é feito o cálculo do ganho de informação (Ganho), que consiste na análise da homogeneidade das subclasses criadas, a citar a métrica Entropia (MITCHELL, 1997). Na Equação 2.7 é ilustrado o cálculo do ganho de informação por meio da Entropia.

$$Ganho = info(T) - \sum_{t=1}^{m} \frac{|T_t|}{|T|} * info(T_t) \tag{2.7}$$

onde

$$info(T) = -\sum_{j=1}^{k} \frac{freq(C_j, T)}{|T|} * \log 2 \left(\frac{freq(C_j, T)}{|T|} \right) \tag{2.8}$$

Sendo $freq(C_j, T)$ o número de amostras T subdivididas no subespaço C_j, T o número total de amostras, k o número de classes existentes, e m o número de subespaços criados na divisão de T.

Random Forest

O *Random Forest* é baseado na construção de várias árvores de decisão em subespaços selecionados aleatoriamente (BREIMAN, 2001). As árvores em diferentes subespaços generalizam sua classificação com intuito de se complementarem, com isso a sua classificação combinada pode ser aprimorada. O critério Gini é utilizado na divisão binária entre os nós da árvore, sendo representado por meio da Equação 2.9.

$$Gini(n) = 1 - \sum_{j=1}^{2} (P_j)^2 \tag{2.9}$$

Sendo *Pj* o número de vezes que a classe j apareceu nas instâncias do conjunto de dados.

AdaBoost

O *AdaBoost* é baseado na técnica de comitês, que combina outros algoritmos com a intenção de construir um resultado mais robusto (FREUND; SCHAPIRE, 1997). Geralmente, para a construção do AdaBoost são utilizadas um conjunto de *Decision Trees*. Neste algoritmo todas as instâncias possuem pesos iguais. No entanto, ao decorrer da execução, os pesos das instâncias classificadas de forma incorreta são incrementados. Ao final, as árvores que tiveram um menor índice de erro são selecionadas para compor o modelo. Assim, a função que representa o novo peso da árvore é definida conforme a Equação 2.10.

$$x = n * log \frac{1 - e}{e} \tag{2.10}$$

Sendo x o novo peso da árvore, n a taxa de aprendizagem (geralmente possui valor 1) e e a soma das amostras classificadas incorretamente.

Logistic Regression

O algoritmo *Logistic Regression* tem como objetivo produzir, a partir de um conjunto de observações, uma estimativa quanto a probabilidade do atributo alvo assumir um determinado valor (PENG; LEE; INGERSOLL, 2002). Para isso, é encontrada uma

função que permita a predição do atributo alvo, frequentemente binário, a partir de um conjunto de atributos contínuos e/ou binários. Considerando um problema de classificação binária, a função que apresenta a probabilidade é definida de acordo com a Equação 2.11.

$$F(i) = \ln\left(\frac{p_i}{1 - p_i}\right) \tag{2.11}$$

Sendo $F(i)$ o logaritmo natural da chance, p_i a probabilidade de ocorrência e $1 - p_i$ a probabilidade da não-ocorrência.

Gradient Boosting

O *Gradient Boosting* também é baseado na técnica de comitês e consiste em um conjunto t de modelos treinados de forma sequencial, no qual, o modelo t tem como objetivo corrigir os erros do modelo *t-1* (CHEN; GUESTRIN, 2016). O erro do modelo na iteração t é definido conforme a Equação 2.12.

$$L_t = l(y_i, \hat{y}_i + f_t(x_i)) \tag{2.12}$$

Sendo *yi* o valor alvo, $\hat{y}i$ a predição obtida pelo modelo t para a instância *xi*, l uma função de erro e n o número total de instâncias.

2.4 Considerações Finais

Mediante a definição dos conceitos envolvidos neste trabalho, foi possível compreender que o processo de Ciência de Dados proporciona a melhoria, transformação e a visualização dos dados, além de possibilitar o desenvolvimento de modelos baseados no aprendizado de máquina. Além disso, em meio aos tipos de aprendizado existentes, o aprendizado supervisionado possibilita a criação de modelos de predição a partir da utilização de algoritmos e um conjunto de dados, sendo utilizados em problemas de classificação e regressão.

3 Mapeamento Sistemático da Literatura

Segundo Kai *et al.* (2008), um Mapeamento Sistemático Literatura (MSL) se caracteriza como um estudo que visa encontrar, avaliar e agregar as contribuições de estudos publicados em relação a uma questão de pesquisa específica. Com isso, neste capítulo é apresentado um MSL com intuito de identificar estratégias preditivas voltadas ao agravamento do quadro clínico de pacientes com a COVID-19.

Por ser um estudo sistemático, o MSL deve seguir um protocolo planejado e fundamentado, de modo a possibilitar que o processo possa ser replicado. Neste MSL, o protocolo utilizado foi baseado no trabalho de Rocha, Nascimento e Nascimento (2018), que estabelece cinco etapas: i) delimitar a questão de pesquisa; ii) dividir a questão de pesquisa em questões específicas; iii) definir as fontes de busca; iv) selecionar os trabalhos relevantes; e, v) avaliar os trabalhos.

3.1 Planejamento

Nesta Seção são especificadas as questões de pesquisa, as fontes de busca e os critérios envolvidos no processo de seleção dos trabalhos.

3.1.1 Questões de Pesquisa

A Questão de Pesquisa (QP) deste MSL se caracteriza como: "Quais as abordagens estão sendo utilizadas na predição do agravamento do quadro clínico de pacientes com a COVID-19?". Para responder este questionamento, foram elaboradas cinco Questões Secundárias (QS):

- **QS1**: Quais algoritmos estão sendo utilizados no desenvolvimento dos modelos?
- **QS2**: Quais recursos tecnológicos (linguagens, ferramentas e bibliotecas) foram utilizadas na análise de dados?
- **QS3**: Qual a origem (fonte dos dados) das bases de dados empregadas na análise de dados?
- **QS4**: Quais variáveis foram consideradas mais relevantes na análise de dados?
- **QS5**: Quais plataformas estão sendo utilizadas na disponibilização dos modelos de predição desenvolvidos?

3.1.2 Processo de Busca

A pesquisa foi direcionada em trabalhos publicados entre 2019 e Setembro de 2021 em seis fontes de busca na área de Ciências da Computação, a citar:

- *ACM Digital Library* (www.dl.acm.org);
- *IEEE Xplore Digital Library* (www.ieeexplore.ieee.org);
- *MedLine* (https://pubmed.ncbi.nlm.nih.gov/);
- MDPI (https://www.mdpi.com/);
- *Scopus* (www.scopus.com);
- *Web of Science* (www.webofknowledge.com).

Para identificar as principais expressões que remetessem aos trabalhos buscados nesta pesquisa, foram definidos termos atrelados as áreas médica e tecnológica. Os termos utilizados deveriam estar no idioma inglês, tendo em vista que as fontes de pesquisa predominantemente indexam artigos científicos escritos em tal idioma. Os termos definidos são apresentados no Quadro 2.

Quadro 2 – Termos de busca.

Área	**Termos**
Médica	COVID; COVID-19; Sars-Cov-2; *Coronavirus*; *Diagnosis.*
Tecnológica	*Machine Learning; Data Mining; Artificial Intelligence; Data Science; Deep Learning*; *Model; Prediction; Forecasting.*

Fonte: Autoria própria.

Para realizar o processo, uma *string* de busca foi definida baseando-se nos termos elencados no Quadro 2, culminando na seguinte *string* de busca: *(COVID OR COVID-19 OR Sars-Cov-2 OR Coronavirus) AND (Machine Learning OR Data Mining OR Artificial Intelligence OR Data Science OR Deep Learning) AND (Model OR Prediction OR Diagnosis OR Forecasting)*. Por conta das especificidades das fontes de busca de cada base foi necessário gerar três *strings* diferentes, mas de mesmo significado:

> **I:** ((“COVID-19” OR “Coronavirus” OR “Covid” OR “Sars-Cov-2”) AND (“Machine Learning” OR “Deep Learning” OR “Data Science” OR “Artificial Intelligence” OR “Data Mining”) AND (“Prediction” OR “Model” OR “Diagnosis” OR “Forecasting”))
>
> **II:** (TI = (Covid OR Coronavirus OR COVID-19 OR Sars-cov-2) AND TI = (Machine Learning OR Data Science OR Artificial Intelligence OR Data Mining OR Deep Learning) AND TI = (Model OR Prediction OR Diagnosis OR Forecasting))

III: ((“Covid” [Title] OR “Coronavirus” [Title] OR “COVID-19” [Title] OR “Sars-cov-2” [Title])) AND (“Machine Learning” [Title] OR “Data Mining” [Title] OR “Artificial Intelligence” [Title] OR “Data Science” [Title] OR “Deep Learning” [Title])) AND (“Model” [Title] OR “Predict” [Title] OR “Diagnosis”[Title] OR “Forecasting”[Title]))

A *string* de busca I foi aplicada nas fontes *ACM*, *IEEE Xplore*, *Scopus* e *MDPI*. Na fonte *Web of Science* foi utilizada a *string* II, enquanto a *string* III foi aplicada na fonte *MedLine*.

3.1.3 Critérios para seleção e avaliação

Para realização da seleção dos trabalhos foi definido um conjunto de Critérios de Inclusão (CI) e Critérios de Exclusão (CE).

Critério de Inclusão:

CI1) Estudos com foco no objetivo da pesquisa.

CI2) Estudos publicados entre os anos de 2019 e 2021 (Setembro).

CI3) Estudos publicados no idioma Inglês.

CI4) Estudos que abordaram pelo menos um modelo estatístico/matemático.

Critérios de Exclusão:

CE1) Estudos repetidos em mais de uma fonte de pesquisa.

CE2) Estudos cuja base de dados utilizada é tipo do tipo não estruturada, tais como: imagens, vídeos e áudio.

CE3) Estudos que não fornecem dados suficientes para responder a nenhuma das questões de pesquisa.

3.2 Seleção dos Trabalhos

O processo de escolha e análise dos artigos foi dividido em quatro estágios. Inicialmente, no Estágio I foi realizada a seleção dos artigos nas fontes de pesquisa, tendo em vista a *string* de busca definida na Seção 3.1.2. Em seguida, foi feita a leitura do título e das palavras-chave dos artigos, juntamente com a avaliação dos critérios de inclusão CI2 e CI3. Ao final, 1245 trabalhos foram selecionados[1].

[1] Trabalhos do Estágio I: https://cutt.ly/ABqcpP6

No Estágio II foram observados, além do título e palavras-chaves, os resumos dos trabalhos, verificando-se também o critério de inclusão CI1 e dois critérios de exclusão (CE1 e CE2). Os trabalhos que atenderam aos critérios de inclusão foram adicionados à Lista de Incluídos (LI1), enquanto os excluídos foram organizados em uma Lista de Excluídos (LE1). Ao final da execução dos critérios, 1087 artigos foram excluídos. Destes, 750 eram duplicados (CE1), 141 utilizaram dados não estruturados (CE2) e 196 não estavam relacionados com o foco desta pesquisa (CI1). Ao término do Estágio II, restaram 158 artigos[2].

No Estágio III foi realizada a leitura das seções de introdução, metodologia e conclusão dos trabalhos presentes LI1. Para este estágio, verificou-se o critério de inclusão (CI4). Com isso, os estudos excluídos foram inseridos na Lista de Excluídos (LE2), enquanto os trabalhos que permaneceram foram adicionados em uma segunda Lista de Incluídos (LI2). Com o término da aplicação dos critérios, 105 trabalhos foram excluídos. Dentre os trabalhos excluídos, 77 não estavam relacionados com o foco desta pesquisa (CI1), 24 fizeram uso de dados não estruturados (CE2) e 4 não forneceram dados suficientes para responder a nenhuma das questões de pesquisa (CE3). Com o término do Estágio III, 53 trabalhos permaneceram na análise[3].

No Estágio IV foram analisados todos os trabalhos da LI2. Neste caso, foi feita uma leitura completa dos trabalhos e uma revisão de todos os critérios de inclusão e exclusão. Ao término deste estágio, os trabalhos que satisfizeram todos os critérios foram direcionados a responder às Questões Secundárias. Ao fim do Estágio IV, 50 trabalhos foram selecionados, de modo a fornecer respostas às QS[4]. Na Figura 3 é apresentada a quantidade de trabalhos incluídos em cada um dos estágios.

3.3 Resultados

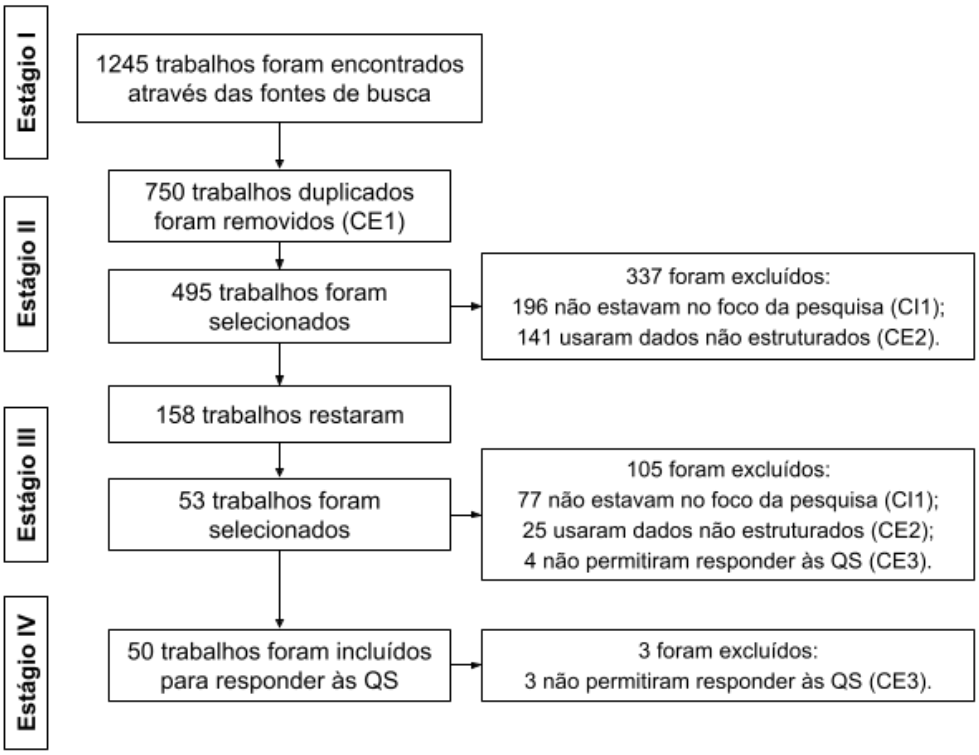
Os trabalhos selecionados foram lidos por completo com intuito de identificar nos resultados alcançados as respostas às Questões Pesquisa definidas na Seção 3.1.1. No Apêndice A é apresentada uma síntese das principais informações dos artigos. A seguir são especificadas as respostas obtidas mediante as Questões Específicas.

3.3.1 QS1: Quais algoritmos estão sendo utilizados no desenvolvimento dos modelos?

Ao analisar o conjunto dos trabalhos selecionados, foi possível catalogar 13 algoritmos que foram utilizados no desenvolvimento dos modelos de predição. A partir da análise

[2] Trabalhos do Estágio II: https://cutt.ly/gBqchkE
[3] Trabalhos do Estágio III: https://cutt.ly/KBqcmMF
[4] Trabalhos do Estágio IV: https://cutt.ly/nBqcY6Q

Figura 3 – Fluxograma dos trabalhos incluídos em cada um dos estágios.



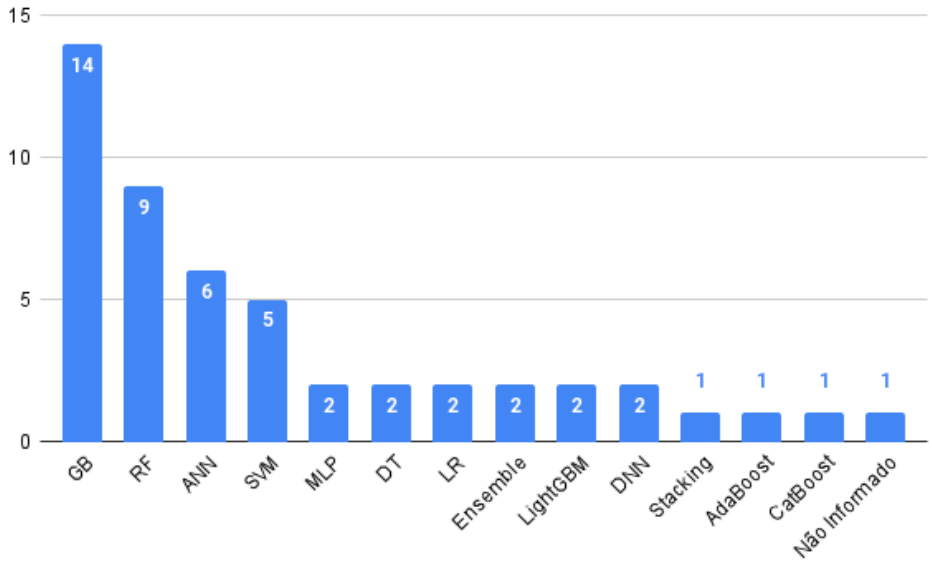
Fonte: Autoria própria.

foi constatado que quatro algoritmos foram utilizados em mais de dois trabalhos, a citar: i) *Gradient Boosting* (GB) presente em 14 trabalhos; ii) *Random Forest* (RF) identificado em nove trabalhos; iii) *Artificial Neural Network* (ANN) presente em seis trabalhos; e, iv) *Support Vector Machine* (SVM) em cinco trabalhos.

Além disso, outros seis algoritmos foram utilizados em pelo menos dois trabalhos, a citar: *Multilayer Perceptron* (MLP), *Decision Tree* (DT), *Logistic Regression* (LR), *Ensemble*, *Light Gradient Boosting Machine* e *Deep Neural Network* (DNN). Em seguida, outros três três algoritmos foram identificados em pelo menos um trabalho, a citar: *Stacking*, *AdaBoost* e *CatBoost*. Por fim, um trabalho não informou o algoritmo utilizado no desenvolvimento do modelo de predição. No gráfico da Figura 4 são apresentados os algoritmos, juntamente com a quantidade de trabalhos que os utilizaram.

Com a organização dos algoritmos utilizados nos trabalhos, é possível perceber algumas similaridades, a citar: i) algoritmos baseados em Árvores de Decisão (como por exemplo: XGB, RF, DT, LightGBM e AdaBoost) foram bastante recorrentes nos trabalhos; ii) variações do algoritmo de Redes Neurais estão sendo utilizados pelos trabalhos (a citar: ANN, MLP e DNN); e, iii) os autores estão utilizando comitês (como por exemplo: *Ensemble*, *Stacking* e *AdaBoost*) com intuito de combinar os algoritmos e obter melhor desempenho preditivo.

Figura 4 – Distribuição dos algoritmos utilizados no desenvolvimento dos modelos de predição.



Fonte: Autoria própria.

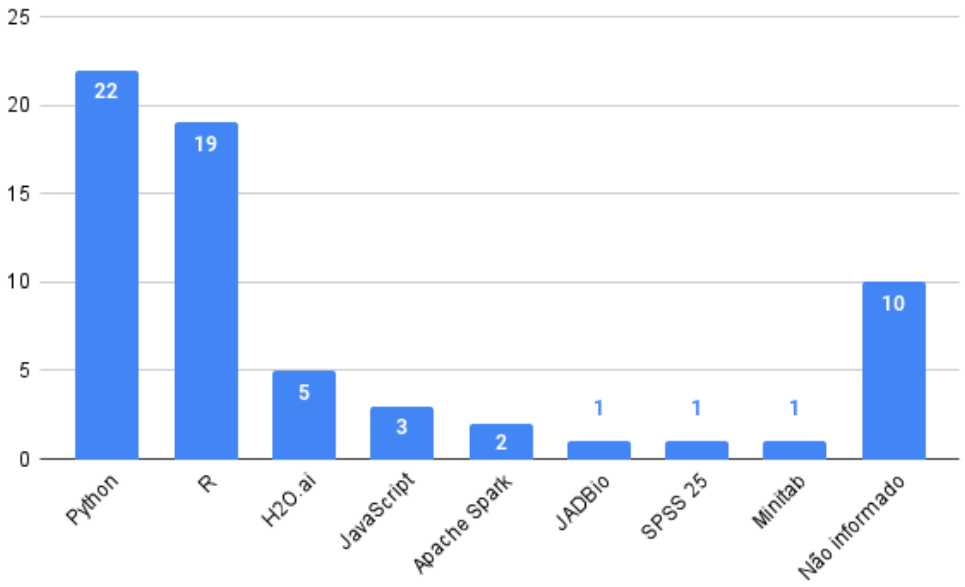
3.3.2 QS2: Quais linguagens de programação e/ou softwares foram utilizadas no desenvolvimento dos modelos de predição?

Baseando-se nos estudos selecionados, foram catalogadas oito linguagens de programação ou softwares que foram utilizados no desenvolvimento dos modelos de predição, a citar: a linguagem Python[5], linguagem R[6], o software H2O.ai[7], a linguagem JavaScipt[8], software Apache Spark [9], o software JADBio[10], o software IBM SPSS[11] e o software Minitab[12]. Na Figura 5 é apresentada a distribuição das linguagens de programação e softwares pela quantidade de vezes empregadas. Vale salientar que o somatório do quantitativo excede o número total de trabalhos, pois em alguns trabalhos os autores fizeram uso de mais de uma linguagem de programação ou software.

Assim, é possível observar que mesmo com a grande variedade de softwares disponíveis para a criação dos modelos de predição, os autores tem utilizado, em sua grande maioria as linguagens de programação, como por exemplo: o Python, R e JavaScript. Esse fato pode ser justificado tendo em vista a quantidade de bibliotecas e pacotes que são disponibilizados para utilização juntamente com as linguagens de programação. Dentre

[5] Python: https://www.python.org/
[6] R: https://www.r-project.org/
[7] H2O.ai: https://h2o.ai/
[8] JavaScript: https://www.javascript.com/
[9] Apache Spark: https://spark.apache.org/
[10] JADBio: https://jadbio.com/
[11] SPSS: https://www.ibm.com/br-pt/spss
[12] Minitab: https://osbsoftware.com.br/produto/spm

Figura 5 – Linguagens de programação e softwares utilizados na criação dos modelos preditivos.



Fonte: Autoria Própria.

alguns dos principais pacotes e bibliotecas identificados nos trabalhos selecionados, podem ser citados: i) a biblioteca Pandas[13], que permite a manipulação e análise de dados através da linguagem Python; ii) a biblioteca Scikit-Learn[14], que possibilita a construção de modelos baseados em aprendizado de máquina por meio da linguagem Python; iii) o *framework* Tidymodels[15], que disponibiliza uma coleção de pacotes para o desenvolvimento de modelos de aprendizado máquina a partir da linaguagem R; e, iv) a biblioteca TensorFlow.js[16], utilizada na criação de modelos baseados em aprendizado de máquina por meio da linguagem JavaScript.

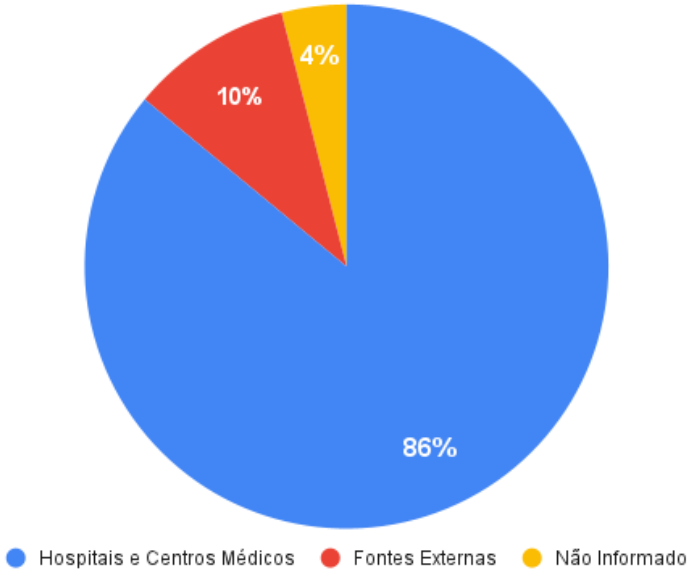
Além disso, notou-se que alguns trabalhos fizeram uso de mais de uma linguagem de programação ou software, com intuito de se beneficiar de seus pacotes e bibliotecas, a mencionar: Python e R (DAS; MISHRA; GOPALAN, 2020; AZNAR-GIMENO *et al.*, 2021; CHEN *et al.*, 2021; SUBUDHI *et al.*, 2021), R e H2O.ai (KIM *et al.*, 2020; HOU *et al.*, 2021; IKEMURA *et al.*, 2021; JAKOB *et al.*, 2021; MAHBOUB *et al.*, 2021), Python e TensorFlow.js (KO *et al.*, 2020), R e TensorFlow (CHUNG *et al.*, 2021), R e Apache Spark (CHENG *et al.*, 2020), SPSS e JASP (ASSAF *et al.*, 2020), R e JADBio , (PAPOUTSOGLOU *et al.*, 2021) e R e Minitab (PAPOUTSOGLOU *et al.*, 2021).

[13] Pandas: https://pandas.pydata.org/
[14] Scikit-Learn: https://scikit-learn.org/stable/index.html
[15] Tidymodels: https://www.tidymodels.org/
[16] TensorFlow.js: https://www.tensorflow.org/js?hl=pt-br

3.3.3 QS3: Qual a origem das bases de dados e quais tipos de dados estão sendo utilizados nos trabalhos?

A partir da análise dos trabalhos selecionados, foi investigado a origem dos dados utilizados no treinamento dos algoritmos de aprendizado de máquina. Dessa forma, assim como ilustrado no gráfico da Figura 6, foi constatado que 86% (43 trabalhos) dos dados foram advindos de parcerias dos respectivos autores com hospitais e centros médicos. Enquanto isso, 10% (5 trabalhos) utilizaram dados de fontes externas, tais como sites governamentais ou comunidades de dados abertos (como por exemplo o Kaggle[17]). Além disso, 4% (2 trabalhos) não inforam a origem dos dados utilizados.

Figura 6 – Origem das bases de dados utilizadas nos trabalhos.

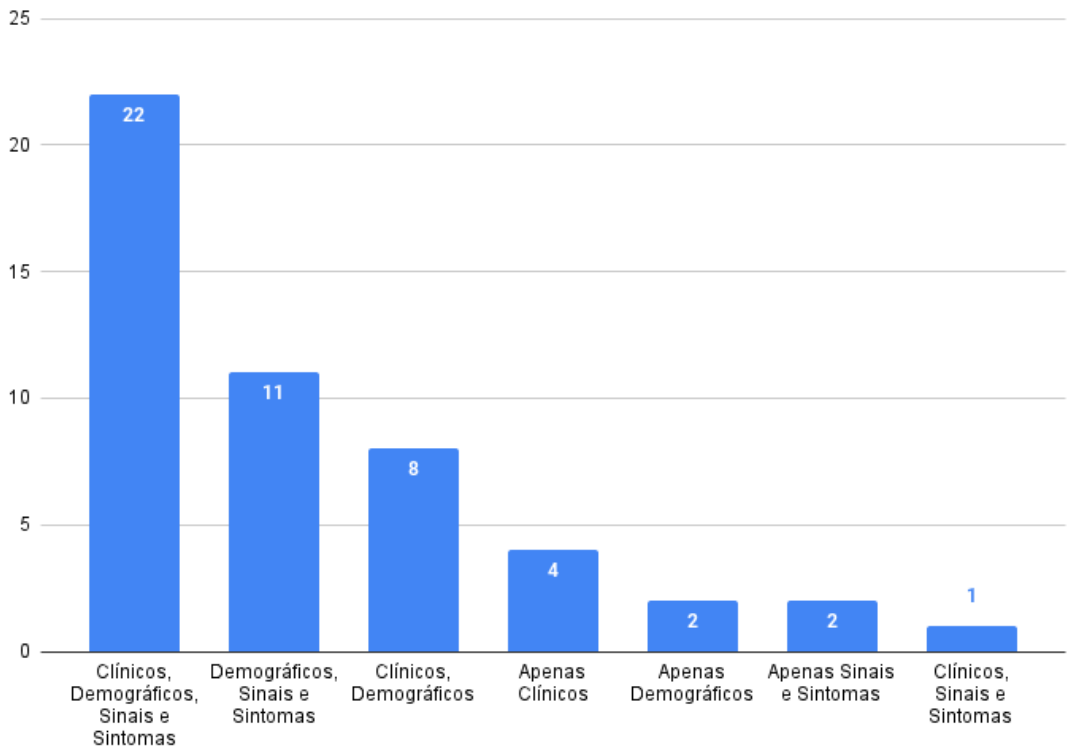


Fonte: Autoria Própria.

Em seguida foi identificado nos trabalhos quais tipos de dados foram utilizados no desenvolvimento dos modelos de predição. Para isso, os dados foram divididos em três categorias: i) clínicos: que são obtidos a partir de exames clínicos; ii) demográficos: informações pessoais do paciente; e, iii) sinais e sintomas: que envolvem a situação atual do paciente em relação aos sintomas e a presença de comorbidades. Assim, a partir o gráfico da Figura 7 é possível observar que: i) 22 trabalhos utilizaram dados clínicos, demográficos e sinais e sintomas; ii) 11 trabalhos utilizaram dados demográficos e sinais e sintomas; iii) oito trabalhos utilizaram dados clínicos e demográficos; iv) quatro trabalhos utilizaram apenas dados clínicos; v) dois trabalhos utilizaram apenas dados demográficos; vi) dois trabalhos utilizaram apenas dados referentes à sinais e sintomas; e, vii) um trabalho utilizou dados clínicos e sinais e sintomas.

[17] Kaggle: https://www.kaggle.com/

Figura 7 – Tipos dos dados utilizados pelos trabalhos.



Fonte: Autoria Própria.

A partir da distribuição dos tipos de dados utilizados nos trabalhos, é possível concluir que a grande maioria dos autores tem feito uso de dados de categorias distintas, onde 44% (22 trabalhos) fizeram uso de dados das três categorias e 84% (42 trabalhos) utilizaram pelo menos dados de duas categorias diferentes, com intuito de avaliar o risco de agravamento do estado clínico dos pacientes em diferentes perspectivas.

3.3.4 QS4: Quais atributos foram consideradas mais relevantes no desenvolvimento dos modelos?

Mediante o desenvolvimento dos modelos preditivos, os autores identificaram os principais fatores relacionados ao agravamento do quadro clínico dos pacientes com a COVID-19. Apesar dos estudos utilizarem um conjunto de atributos diferentes para a condução da análise e elaboração do modelo preditivo, foi possível constatar similaridades entre as características descobertas. Para ilustrar os fatores atrelados a possibilidade de agravamento, foram selecionados os três atributos de maior importância em cada estudo. Na Tabela 1 são apresentados os três principais atributos que aparecerem em pelo menos dois trabalhos.

Devido a variabilidade dos dados utilizados no desenvolvimento dos estudos seleci-

Tabela 1 – Principais atributos relacionados ao agravamento de pacientes com a COVID-19 agrupados por estudo.

Atributo	**Natureza**	**Trabalhos**
Idade	Demográfico	22
Proteína C	Clínico	12
Linfócitos	Clínico	8
Neutrófilos	Clínico	8
Saturação O2	Sinais e Sintomas	7
Sexo	Demográfico	5
Lactato Desidrogenase	Clínico	4
Frequência Respitatória	Sinais e Sintomas	4
IMC	Demográfico	3
Problema Renal	Sinais e Sintomas	3
Cálcio no Sangue	Clínico	3
Nitrogênio Ureico	Clínico	3
Sódio	Clínico	2
Creatinina	Clínico	2
Dispneia	Sinais e Sintomas	2
Temperatura	Sinais e Sintomas	2
Procalcitonina	Clínico	2
Pressão Arterial	Sinais e Sintomas	2

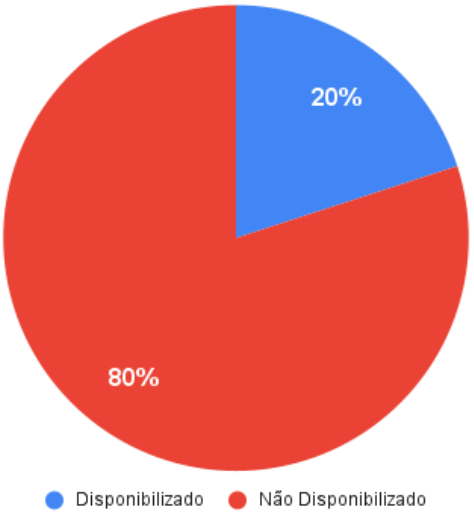
Fonte: Autoria própria.

onados, a maioria das variáveis tiveram aparição em poucos estudos. No entanto, também é possível identificar que: i) a idade foi um atributo destacado em 22 (44%) dos trabalhos, fato que reforça a importância desse atributo mesmo em trabalhos que utilizaram bases de dados distintas; ii) a proteína C, quantidade de linfócitos e neutrófilos, foram os principais atributos clínicos relatadas nos trabalhos; e, iii) atributos relacionados ao sintoma de dificuldade respiratória foram identificados com certa frequência nos trabalhos, tais como: saturação de O2, frequência respiratória e dispneia.

3.3.5 QS5: Quais plataformas estão sendo utilizadas na disponibilização dos modelos de predição desenvolvidos?

Por fim, foi investigado quantos trabalhos disponibilizaram os modelos de predição após o seu desenvolvimento. Como observado no gráfico da Figura 8, é possível constatar que dentre os 50 modelos de predição criados, apenas 10 (20%) foram disponibilizados para utilização, enquanto 40 (80%) não foram disponibilizado. Já em relação as plataformas utilizadas para disponibilização dos modelos de predição, um modelo foi disponibilizado na plataforma de dispositivos móveis *Android*, enquanto os outros nove foram disponibilizados através de *websites*. O link de acesso aos modelos pode ser consultado no Apêndice A.

Figura 8 – Quantos modelos de predição foram disponibilizados?



Fonte: Autoria Própria.

3.4 Considerações Finais

Este MSL auxiliou na identificação do cenário atual de trabalhos científicos sobre o desenvolvimento de soluções preditivas relacionadas ao agravamento do quadro clínico de pacientes com a COVID-19. Por meio dos resultados apresentados foi constatado que: *i)* um conjunto de algoritmos de aprendizado de máquinas vêm sendo utilizadas no desenvolvimento dos modelos de predição, a citar, por exemplo: *Gradient Boosting*, *Random Forest*, *Artificial Neural Network* e *Support Vector Machine*; *ii)* as linguagens de programação Python e R estão sendo utilizadas com frequência na análise dados e no desenvolvimento dos modelos preditivos, principalmente devido a grande variedade de pacotes e bibliotecas disponíveis; *iii)* além das linguagens de programação, alguns *softwares* estão sendo empregados tanto na análise e visualização dos dados de forma gráfica, quanto na automatização das etapas que envolvem a criação dos modelos; *iv)* mesmo com a disponibilidade de bases de dados abertas com informações relacionadas aos casos de COVID-19, notou-se que a maioria dos trabalhos selecionados utilizaram bases obtidas diretamente de hospitais e centros médicos, com o objetivo de coletar e avaliar uma maior variedade de informações de uma região geográfica específica; *v)* em relação aos principais atributos atrelados ao agravamento do quadro clínico, a idade do paciente, a presença de sintomas relacionados a dificuldade respiratória e os indicadores de proteína C, linfócitos e neutrófilos, foram identificados em uma parcela dos trabalhos selecionados; e, *vi)* por fim, mesmo com o desenvolvimento dos modelos de predição, foi constatado que poucos modelos tem sido disponibilizados, o que dificulta a sua utilização.

4 Base de Dados e Pré-Processamento

Neste capítulo são apresentadas o conjunto de etapas necessárias para a concepção do modelo de predição (Seção 4.1). Na sequência é apresentada a base de dados utilizada nesta pesquisa (Seção 4.2). Em seguida são apesentados os procedimentos utilizados no pré-processamento (Seção 4.3). Mais adiante é apresentada uma análise nos dados dos pacientes internados e que vieram a óbito, com intuito de identificar similaridades (Seção 4.4). Logo após são descritas as novas bases geradas após a realização das etapas (Seção 4.5). Por fim, são apresentadas as considerações finais do capítulo (Seção 4.6).

4.1 Etapas da Concepção do Modelo de Predição

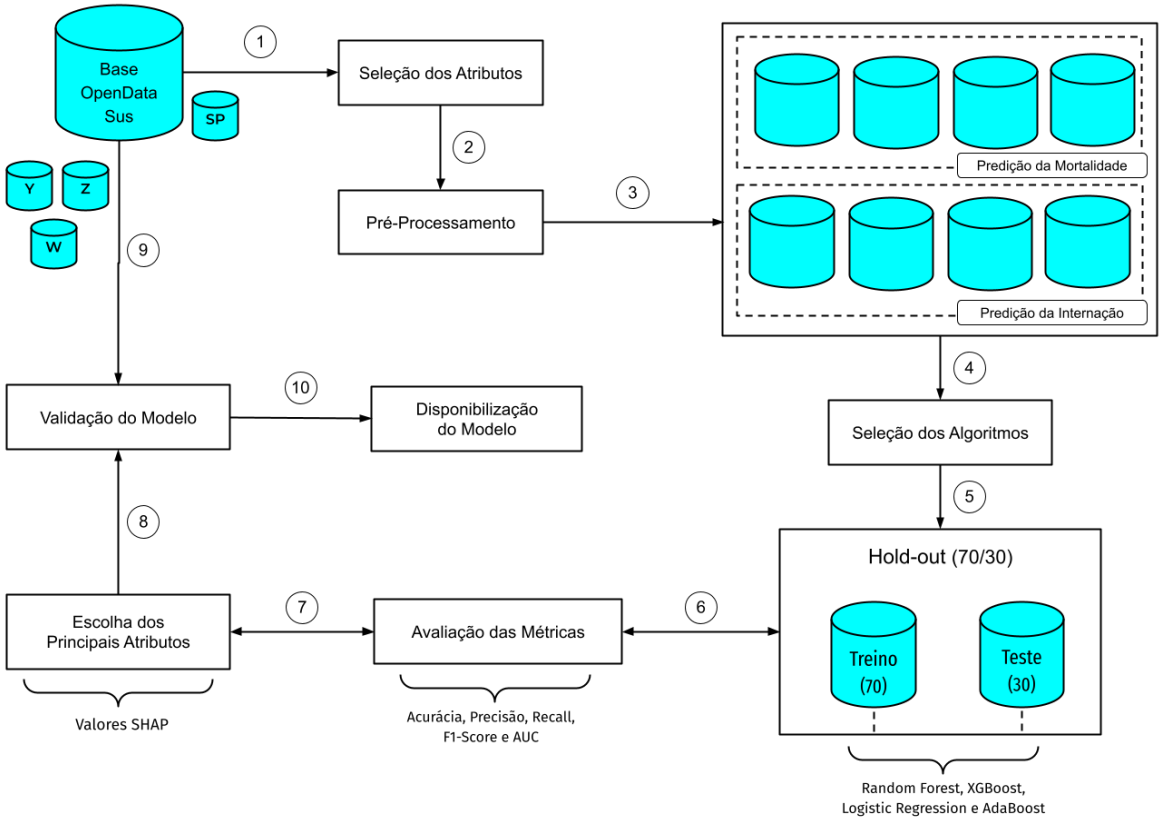
Para a concepção do modelo de predição foram definidas as seguintes etapas: *i)* escolha da base de dados; *ii)* seleção dos atributos; *iii)* pré-processamento dos dados; *iv)* definição de novas bases de dados; *v)* seleção dos algoritmos; *vi)* treinamento e teste dos modelos; *vii)* avaliação dos modelos a partir das métricas de desempenho; *viii)* avaliação dos principais atributos; *ix)* validação dos modelos criados, e; *x)* disponibilização dos modelos em uma aplicação *Web*, conforme ilustrado na Figura 9.

Para a realização deste trabalho foram utilizados os dados da plataforma OpenDataSus, que contém informações sobre notificações de casos de COVID-19 nos estados brasileiros. Dentre as bases de dados disponíveis, a base de dados referente ao estado de São Paulo (SP) foi escolhida para o treinamento dos modelos por possuir um maior número de instâncias. Em seguida, foram selecionados os atributos utilizados como variáveis preditoras para ambos modelos, de internação e mortalidade. Entre estes, dados dos seguintes tipos: demográficos, histórico de vacinação, sintomas, comorbidades e evolução do estado de saúde dos pacientes.

Com intuito de construir os dois modelos de agravamento, o primeiro levando em conta a probabilidade de internação e o segundo a probabilidade de mortalidade, a base de dados foi dividida em dois conjuntos de dados, um para cada modelo. No entanto, diante do desbalanceamento dos dados, foram criadas versões balanceadas de cada um dos conjuntos. Com isso, foi utilizada a técnica SMOTE (*do inglês, Synthetic Minority Over-sampling Technique*), que permite gerar instâncias das classes minoritárias a partir dos seus vizinhos mais próximos (CHAWLA *et al.*, 2002).

Em seguida, para realizar a etapa de seleção dos algoritmos, foi selecionada a biblioteca *PyCaret*, baseada no aprendizado de máquina automatizado e que possibilita a avaliação de desempenho de 14 algoritmos pré-selecionados a partir de um conjunto de

Figura 9 – Descrição das etapas de concepção do modelo de predição.



Fonte: Autoria Própria.

métricas(ALI, 2020). Nesse sentido, a partir dos resultados obtidos ao utilizar a biblioteca, foi observado que os algoritmos *AdaBoost*, *Logistic Regression*, *Random Forest* e *Gradient Boosting* obtiveram os melhores resultados, tendo em vista as métricas de Acurárica, Precisão, *Recall* e *F1-Score*.

Após isso, foi utilizada a técnica *hold-out* para a realização do treinamento e teste dos quatro algoritmos selecionados. Para isso, foi utilizada uma proporção de 70/30, sendo 70% das instâncias para treinamento e 30% para teste. Na etapa seguinte, visando avaliar os resultados alcançados pelos algoritmos, foram utilizadas as métricas: Acurácia, Precisão, *F1-Score*, *Recall* e AUC (do inglês, *Area Under Curve*).

Para a avaliação dos atributos relevantes, ou seja, os atributos que estatisticamente alcançaram o maior impacto na predição, foi utilizada a técnica SHAP (do inglês, *Shapley Additive exPlanations*), que permite identificar as contribuições de cada um dos atributos em relação a predição do modelo (LUNDBERG; LEE, 2017). Após isso, com objetivo de avaliar os modelos desenvolvidos, foram utilizadas as bases de dados com informações dos outros estados brasileiros do OpenDataSus. Por meio dessa validação, foi possível verificar se os modelos alcançaram resultados similares aos obtidos utilizando a base de dados do estado de São Paulo.

Por fim, visando demonstrar como os modelos propostos seriam utilizados na prática, uma aplicação *Web* foi desenvolvida. O código-fonte de toda a análise de dados apresentada neste trabalho encontra-se disponível em um repositório no GitHub (HOLANDA, 2022). Nas subseções a seguir, serão descritas em mais detalhes as etapas da metodologia utilizada neste trabalho.

4.2 Base de Dados

Os dados utilizados neste trabalho foram coletados no dia 4 de abril de 2022 por meio do OpenDataSus, um sistema web mantido pelo Ministério da Saúde do Brasil que disponibiliza informações sobre notificações de casos de COVID-19. O OpenDataSus entrou em vigor em março de 2020, e desde então, periodicamente, os dados são atualizados.

Diante do quantitativo de instâncias, os dados são divididos em um conjunto de lotes, que são organizados por cada um dos 27 estados do Brasil. Na Figura 10 é ilustrada a distribuição do número de instância em relação aos estados brasileiro, enquanto na Tabela 2 são apresentados o número total de instâncias em cada estado. Dessa forma, por ser a base de dados com o maior quantitativo de instâncias, a base de dados do estado de São Paulo (SP) foi escolhida para ser utilizada no treinamento e teste dos modelos, enquanto as bases de dados restantes foram utilizadas na etapa de validação do modelo desenvolvido.

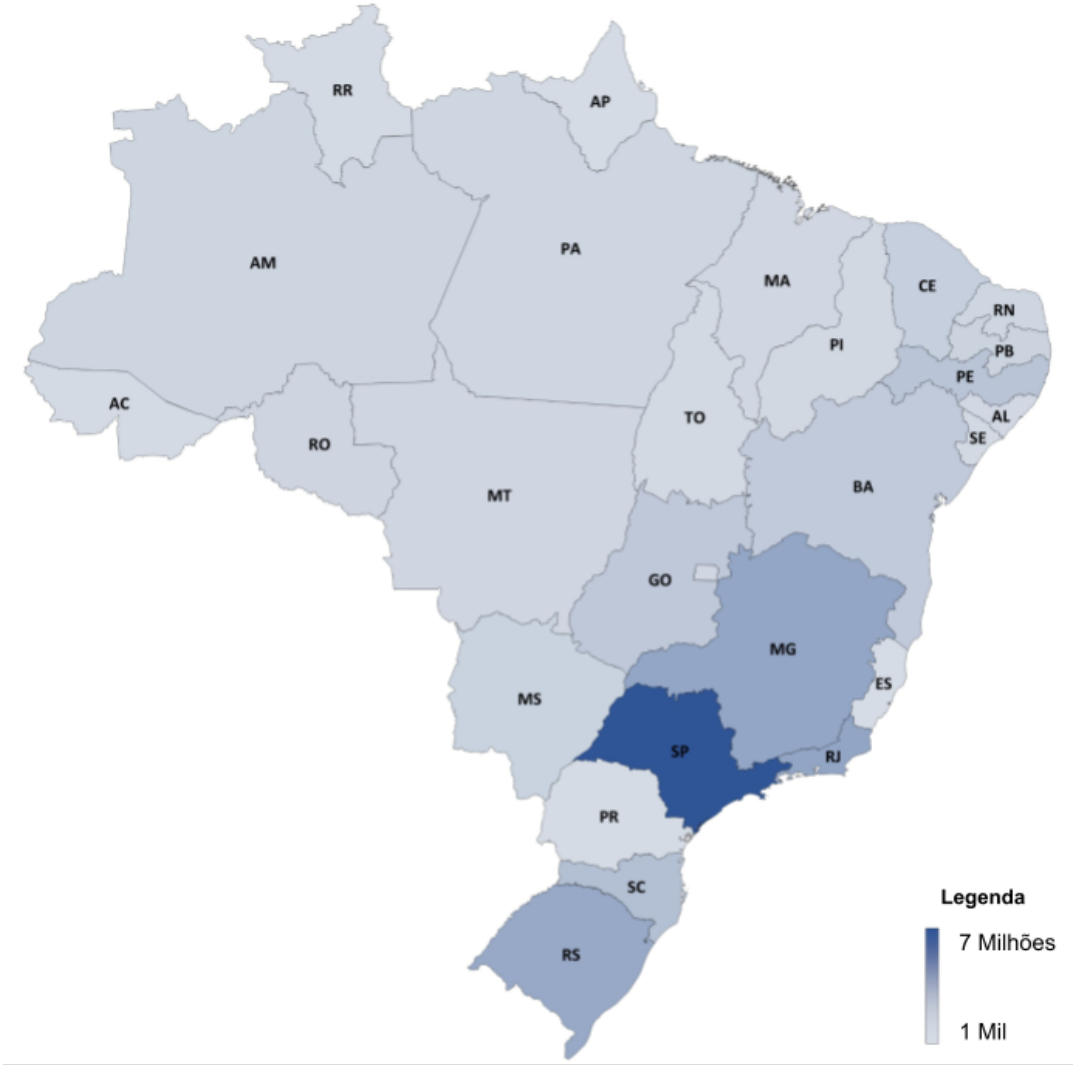
Tabela 2 – Descrição do quantitativo de instâncias das bases de dados do OpenDataSus por estados.

Base (UF)	Instâncias	Base (UF)	Instâncias
SP	6965418	PA	314237
RJ	2835211	MT	295676
MG	2772030	MA	252809
RS	2621814	DF	223554
SC	1452800	PI	201964
PE	1192621	AL	188038
GO	1013164	SE	163819
BA	923350	TO	140349
CE	683198	RR	105420
MS	518009	AC	105014
PB	470476	AP	49639
AM	350349	PR	44962
RO	344060	ES	16285
RN	323365		

Fonte: Autoria própria.

Cada uma das bases de dados dos estados são compostas por 64 atributos, sendo catalogados de acordo os campos da Ficha de Notificação de Casos Suspeitos de COVID-19, utilizada pelo Ministério da Saúde (Anexo A). Os dados coletados dos pacientes são

Figura 10 – Distribuição do número de instâncias pelos estados do Brasil.



Fonte: Autoria própria.

divididos em cinco principais seções: *i)* dados demográficos; *ii)* histórico de vacinação; *iii)* sintomas e comorbidades; *iv)* informações sobre o teste de COVID-19, e *v)* evolução do estado de saúde do paciente.

4.3 Pré-Processamento

Para a realização do pré-processamento dos dados, foi utilizada a linguagem de programação Python e o ambiente de desenvolvimento Jupyter Notebook (KLUYVER *et al.*, 2016). Além disso, foi empregada a biblioteca Pandas, que permite a manipulação e análise de dados a partir de um conjunto de funções e operações previamente definidas (MCKINNEY, 2010). Em seguida, dentre os 64 atributos disponíveis na base de dados, 11 atributos foram selecionados, a citar: *dataNotificacao*, dataInicioSintomas, *sexo*, *racaCor*, *idade*, *sintomas*, *condicoes*, *dataPrimeiraDose*, *dataSegundaDose*, *classificacaoFinal* e

evolucaoCaso.

O atributo *classificacaoFinal* destaca o resultado do diagnóstico (positivo ou negativo) de COVID-19. Neste trabalho foram selecionados apenas os pacientes que foram confirmados com a doença. Assim, os pacientes que não foram confirmados com a COVID-19 foram removidos. Posteriormente, de modo a organizar e analisar os pacientes em grupos, o atributo idade foi categorizado. A categorização foi baseada na escala definida pelo último censo demográfico realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE, 2011), resultando em um total de oito faixas etárias.

Após isso, foram removidas as instâncias que possuíam valores ausentes nos atributos de *sexo* e *racaCor*. O atributo *sexo* foi categorizado em dois valores: feminino e masculino, já o atributo *racaCor* foi categorizado em cinco valores: branca, preta, amarela, parda ou indígena. É válido destacar que categorização de ambos os atributos foi baseada nos campos existentes da Ficha de Notificação de Casos Suspeitos de COVID-19, destacada na Seção 4.2.

Por meio dos atributos *dataPrimeiraDose* e *dataSegundaDose* foi possível identificar em qual data o paciente recebeu a primeira e segunda dose da vacina contra a COVID-19, respectivamente. Caso o atributo estivesse vazio, significa que o paciente não recebeu a respectiva dose. Diante disso, foi feita a conversão dos dois atributos em um novo (*qntVacinas*), que destaca o número de doses tomadas pelo paciente. No Quadro 3 são descritos os atributos demográficos e histórico de vacinação.

Por meio do atributo *sintomas* foi possível identificar a listagem dos sintomas destacados pelo paciente. Diante dessa característica, foi realizada a conversão do atributo, que se encontrava em formato textual, em um conjunto de atributos categóricos, destacando a presença ou ausência de cada sintoma. Assim, o atributo foi divido em: *assintomatico*, *febre*, *dorDeGarganta*, *dispneia*, *tosse*, *coriza*, *dorDeCabeca*, *disturbiosGustatorios* e *disturbiosOlfativos*. No Quadro 4 são descritos os atributos relacionados aos sintomas.

Através do atributo *condicoes* foi possível identificar as comorbidades destacadas por cada paciente. De forma similar aos sintomas, o atributo *condicoes* também foi categorizado, resultando em nove comorbidades: *diabetes*, *obesidade*, *renal*, *respiratoria*, *imunossupressao*, *fragilidadeImuno*, *gestante*, *cardiaca* e *puerpera*. No Quadro 5 são descritos os atributos relacionados às comorbidades.

O atributo *evolucaoCaso* destaca a evolução do estado de saúde do paciente, sendo a variável alvo deste estudo. Assim, instâncias que possuíam valores ausentes foram removidas. Em relação aos possíveis valores do atributo, foram definidas três classificações: *i) Óbito*, caso o paciente tenha morrido; *ii) Internado*, caso o paciente estivesse internado em UTI, e *iii) Curado*, caso o paciente estivesse curado ou em tratamento domiciliar. Com isso, ao final do pré-processamento, a base de dados continha um total de 879.832

Quadro 3 – Descrição dos atributos demográficos e histórico de vacinação.

Atributo	**Descrição**	**Valor**	**Significado**
sexo	Sexo do Paciente	0	Mulher
		1	Homem
racaCor	Raça do Paciente	0	Branca
		1	Preta
		2	Amarela
		3	Parda
		4	Indígena
faixaEtaria	Faixa-etária do Paciente	0	Entre 0 e 11 anos
		1	Entre 12 e 17 anos
		2	Entre 18 e 29 anos
		3	Entre 30 e 44 anos
		4	Entre 35 e 59 anos
		5	Entre 60 e 74 anos
		6	Entre 75 e 89 anos
		7	Mais de 89 anos
qntVacinas	Doses de Vacinas tomadas	0	Nenhuma dose
		1	Uma dose
		2	Duas doses

Fonte: Autoria própria.

instâncias e 24 atributos.

4.4 Análise dos Dados

As análises apresentadas nesta seção foram feitas com foco na observação dos pacientes internados e que vieram a óbito em relação a faixa etária, sintomas, comorbidades e número de vacinas tomadas. Na Figura 11 é apresentada a relação entre o número de internações e óbitos, tendo em vista a faixa etária.

Como observado na Figura 11a, é possível identificar que a faixa etária com maior número de internações é a representada por pacientes idosos, de 60 à 74 anos. Além disso, as faixas etárias de 30 à 44 anos, 45 à 59 anos e 75 à 89 anos apresentaram um quantitativo de pacientes internados similar, o que pode representar uma tendência de internação em pacientes com a idade mais avançada. Por fim, outro fato a ser destacado é o elevado número de internações de crianças de até 11 anos.

Além disso, a partir da Figura 11b é observado uma tendência no quantitativo de óbitos conforme são apresentadas faixas etárias com idades mais avançadas, o que pode reforçar que a idade avançada pode ser um potencial fator de agravamento em relação as chances de mortalidade. De forma similar a Figura 11a, as duas faixas etárias com maior número de óbitos foram as de 60 à 74 anos e 75 à 89 anos. Outro fato que vale

Quadro 4 – Descrição dos atributos relacionados aos sintomas.

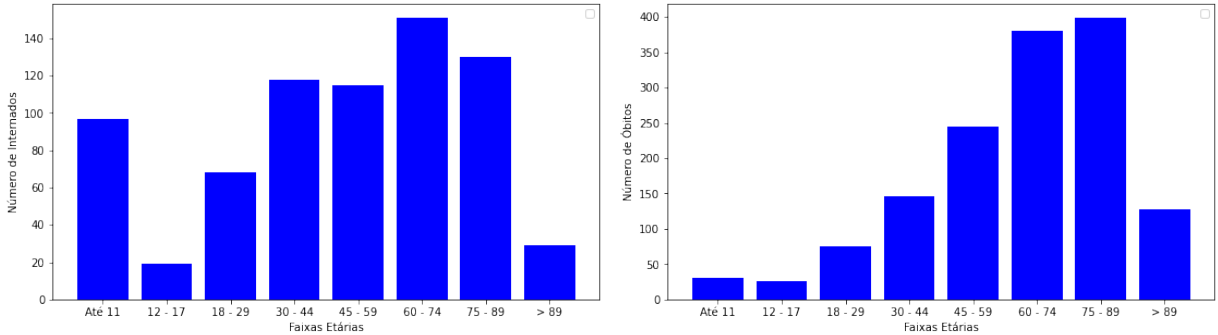
Atributo	**Descrição**	**Valor**	**Significado**
faixaSintomas	Dias desde o primeiro sintoma	0 1 2 3 4 5 6	Entre 0 e 3 dias Entre 4 e 6 dias Entre 7 e 9 dias Entre 10 e 12 dias Entre 13 e 15 dias Entre 16 e 18 dias Mais de 18 dias
assintomatico	Paciente assintomático	0 1	Não Sim
febre	Apresentou febre	0 1	Não Sim
dorDeGarganta	Apresentou dor de garganta	0 1	Não Sim
dispneia	Apresentou falta de ar	0 1	Não Sim
tosse	Apresentou tosse	0 1	Não Sim
coriza	Apresentou coriza	0 1	Não Sim
dorDeCabeca	Apresentou dor de cabeça	0 1	Não Sim
disturbiosGustatorios	Apresentou perda do paladar	0 1	Não Sim
disturbiosOlfativos	Apresentou perda do olfato	0 1	Não Sim

Fonte: Autoria própria.

ser salientado é que mesmo com um elevado número de crianças de até 11 anos sendo internadas, o quantitativo de óbitos dessa faixa etária é relativamente baixo, se comparado com as demais.

Figura 11 – Análise do número de internações e óbitos em relação a faixa etária.

(a) Número de internados em relação a faixa etária. (b) Número de óbitos em relação a faixa etária.



Fonte: Autoria Própria.

Quadro 5 – Descrição dos atributos relacionados às comorbidades

Atributo	**Descrição**	**Valor**	**Significado**
diabetes	Possui diabetes	0	Não
		1	Sim
obesidade	Possui algum grau de obesidade	0	Não
		1	Sim
renal	Possui algum problema renal	0	Não
		1	Sim
respiratoria	Possui algum problema respiratório	0	Não
		1	Sim
imunossupressao	Possui Doença Autoimune	0	Não
		1	Sim
fragilidadeImuno	Possui Imunidade baixa	0	Não
		1	Sim
gestante	É gestante	0	Não
		1	Sim
cardiaca	Paciente possui algum problema cardíaco	0	Não
		1	Sim
puerpera	Está em período de pós-parto	0	Não
		1	Sim

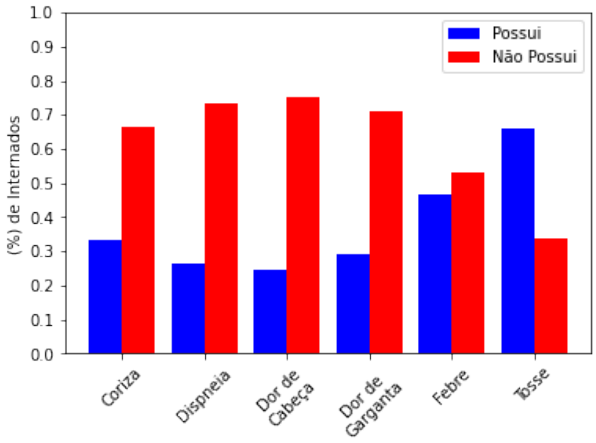
Fonte: Autoria própria.

Em seguida, na Figura 12 é apresentada a relação entre a porcentagem de internações e óbitos tendo em vista os sintomas relatados pelos pacientes. A partir da Figura 12a é possível observar que o sintoma de tosse foi o único onde a porcentagem dos pacientes internados que manifestaram o sintoma foi maior do que os pacientes que não manifestaram. Em relação ao sintoma de febre, é possível notar uma proporção quase igualitária entre a parcela de pacientes internados que manifestaram o sintoma e parcela que não manifestou. Já em relação aos sintomas de coriza, dispneia, dor de cabeça e dor de garganta, foi notado uma semelhança na proporção dos dados, onde cerca de 70% dos pacientes internados não manifestaram os sintomas, enquanto cerca de 30% manifestaram. Nesta análise não foram utilizados os atributos *disturbiusGustatorios* e *disturbiosOlfativos*, pois, ao avaliar os registros da base de dados, nenhum paciente havia manifestado algum dos distúrbios.

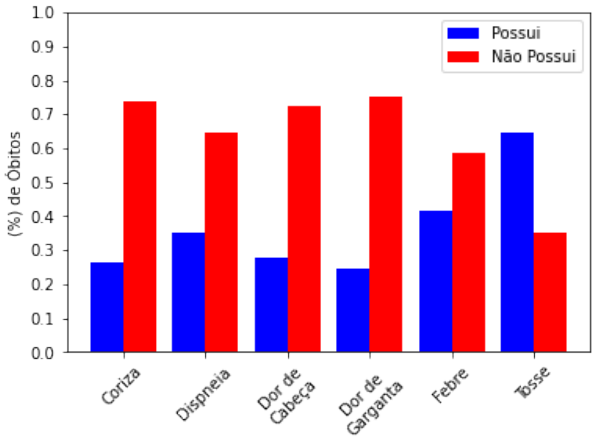
Na sequência, na Figura 12b é observado que a tosse foi a única dentre os sintomas que a porcentagem óbitos que manifestaram o sintoma foi maior que a porcentagem que não manifestou. Além disso, assim como na Figura 12a, a parcela de óbitos que relatou tosse foi próxima da parcela que não relatou. Os sintomas de coriza, dor de cabeça e dor de garganta obtiveram porcentagens similares, totalizando pouco menos de 30% dos óbitos que manifestaram os sintomas. No entanto, se comparado com os pacientes internados, é possível verificar que os pacientes que foram a óbito relataram uma porcentagem ligeiramente maior de falta de ar (dispneia).

Figura 12 – Análise do número de internações e óbitos em relação aos sintomas.

(a) Número de internados em relação aos sintomas.



(b) Número de óbitos em relação aos sintomas.

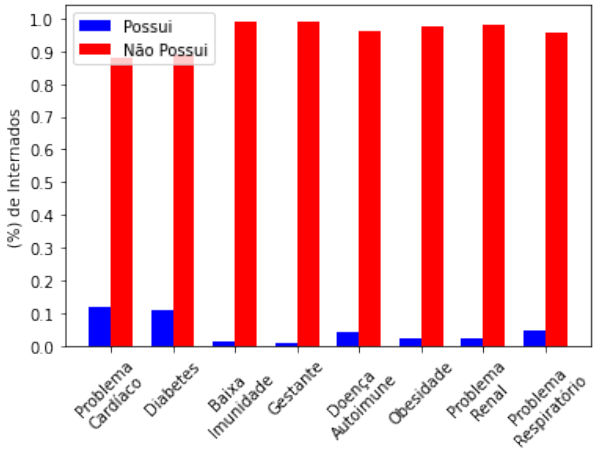


Fonte: Autoria Própria.

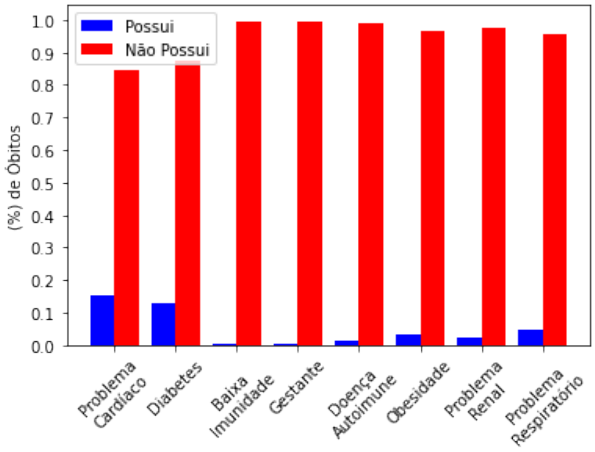
Posteriormente, foi analisada a relação entre a porcentagem de internações e óbitos em relação a presença de comorbidades, conforme ilustrado na Figura 13. Ao verificar as Figuras 13a e 13b, é possível identificar uma forte similaridade, onde a maioria das parcelas dos casos, sejam de internação ou óbito não possuíam comorbidades. No entanto, a presença de algum problema cardíaco, diabetes e problema respiratório, foram as comorbidades de maior proporção dentre as demais. Além disso, é observado um ligeiro aumento das respectivas três comorbidades citadas anteriormente na Figura 13b, se comparado com a Figura 13a.

Figura 13 – Análise do número de internações e óbitos em relação a presença de comorbidades.

(a) Número de internados em relação a presença de comorbidades.



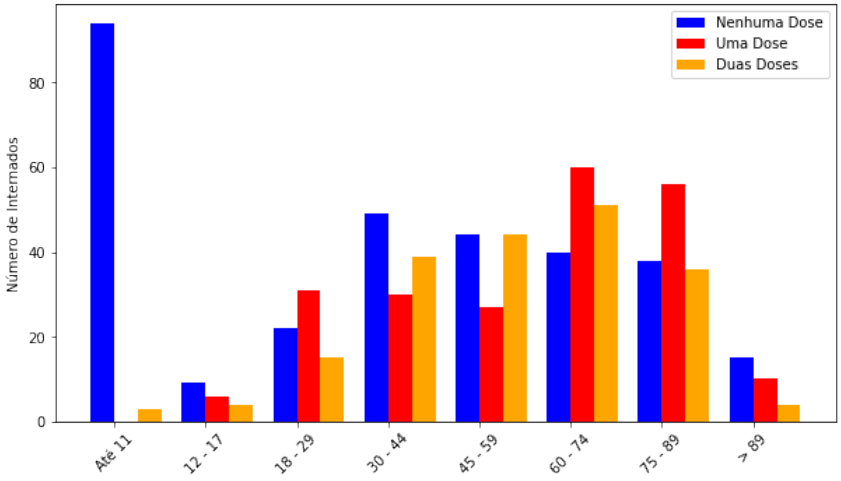
(b) Número de óbitos em relação a presença de comorbidades.



Fonte: Autoria Própria.

Ao final, foi analisada a relação entre a porcentagem de internações e óbitos em relação ao número de doses de vacina, conforme ilustrado nas Figuras 14 e 15, respectivamente. A partir da Figura 14 é observado que o grande número de crianças de até 11 anos internadas não estavam vacinadas. Além disso, nas faixas etárias de 12 à 17 anos, 30 à 44 anos e idade maior que 89 anos, é observado que a maioria dos pacientes internados também não estavam vacinados. No entanto, nas faixas etárias de 18 à 29 anos, 60 à 74 anos e 75 à 89 anos, o quantitativo de internações é maior em pacientes que tomaram pelo menos uma dose. Outro fato que vale a pena ser ressaltado é que em nenhuma faixa etária o número de internados com duas doses é maior que o número de internações com uma ou nenhuma dose.

Figura 14 – Número de internados em relação ao número de doses da vacina.



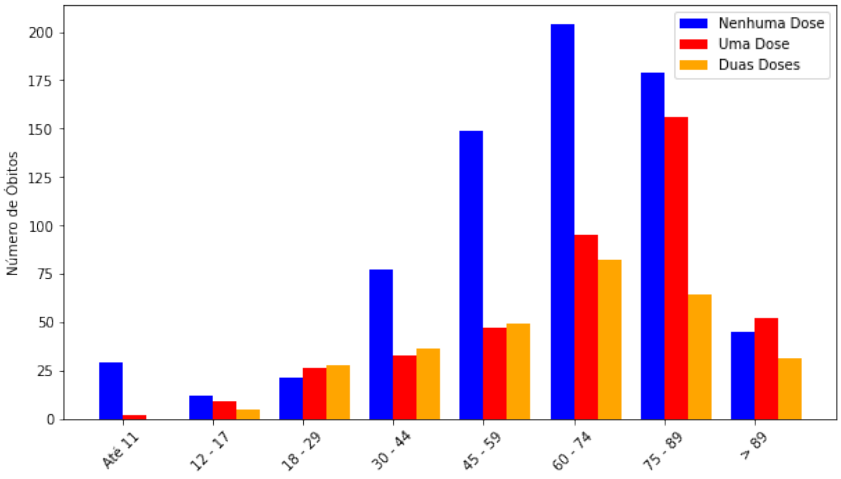
Fonte: Autoria Própria.

Em seguida, na Figura 15 é observado que na maioria das faixas etárias (até 11 anos, 12 à 17, 30 à 40 anos, 45 à 59 anos, 60 à 74 anos e 75 à 89 anos) o número óbitos é maior em pacientes que não foram vacinados. No entanto, na faixa etária de 18 à 29 anos o número de óbitos é ligeiramente maior em pacientes que foram vacinados com duas doses. Outro fato que vale ser destacado é que mesmo os idosos e adultos sendo as faixas etárias priorizadas na vacinação, em linhas gerais, a maior proporção de óbitos nessas faixas etárias foram de pacientes que não estavam vacinados.

4.5 Novas Bases Geradas

Como observado na Seção 4.3, ao finalizar da etapa de pré-processamento, a base de dados estava organizada em um total de 879.832 instâncias e 24 atributos. Tendo em vista o objetivo do trabalho de desenvolver dois modelos de predição (mortalidade e internação), a base pré-processada foi dividida em duas partes, a primeira contendo apenas

Figura 15 – Número de óbitos em relação ao número de doses da vacina.



Fonte: Autoria Própria.

os pacientes que vieram a óbito ou que foram curados, enquanto a segunda continha dados dos pacientes que foram internados ou curados.

No entanto, ao avaliar a distribuição dos valores em relação à variável alvo (*evolucaoCaso*), foi observado que as classes estavam desbalanceadas. No qual, 877.674 instâncias eram classificados como *Curado*, 1.431 como *Óbito* e 727 como *Internado*. Diante deste contexto, foi realizado o balanceamento dos dados como forma de minimizar o problema.

Com intuito de testar algumas variações dos dados, foram geradas quatro versões distintas para serem utilizadas em cada um dos dois modelos. As duas primeiras versões foram desenvolvidas utilizando a redução do número de elementos da classe majoritária. Enquanto para criação das últimas duas versões foi utilizada a biblioteca SMOTE, baseada na técnica estatística de mesmo nome, que possibilita balancear o número de elementos da classe minoritária com a majoritária (LEMAITRE; NOGUEIRA; ARIDAS, 2017).

A SMOTE utiliza todo o conjunto de dados como entrada, mas aumenta apenas o número dos elementos minoritários. Para isso, são geradas novas instâncias com base nas características dos seus vizinhos mais próximos, obtendo assim, um número maior de elementos da classe minoritária e equilibrando as classes.

Em relação ao modelo de predição de mortalidade, a primeira base de dados (Base A) foi organizada em uma proporção de 60/40, onde 60% (2.146) dos dados eram pacientes curados e 40% (1.431) de pacientes que vieram a óbito. Enquanto isso, a segunda base de dados (Base B) continha uma proporção de 70/30, sendo 70% (3.339) curados e 30% (1.431) que vieram a óbito.

A terceira base de dados (Base C) foi gerada a partir da técnica SMOTE e utilizou a primeira versão da base de dados para equilibrar as instâncias em uma proporção de 50/50, sendo 50% (2.146) distribuídos em cada uma das classes. Por fim, a quarta versão (Base D)

foi criada de forma similar a terceira, mas utilizando a segunda versão da base de dados para equilibrar o número de instâncias em uma proporção de 50/50, onde 50% (3.339) dos dados foram divididos nas duas classes. Na Tabela 3 são descritas as informações sobre as quatro versões das bases de dados a serem utilizadas na concepção do modelo preditivo do risco de mortalidade.

Tabela 3 – Descrição das versões balanceadas das bases de dados a serem utilizadas na construção do modelo de predição da mortalidade.

Identificador	**Instâncias**	**Curado**	**Óbito**
Base A	3577 (100%)	2146 (60%)	1431 (40%)
Base B	4770 (100%)	3339 (70%)	1431 (30%)
Base C	4292 (100%)	2146 (50%)	2146 (50%)
Base D	6678 (100%)	3339 (50%)	3339 (50%)

Fonte: Autoria própria.

De forma similar ao modelo anterior, foram geradas quatro versões das bases de dados para o modelo de predição do risco de internação. A primeira base de dados (Base A') foi organizada em uma proporção de 60/40, onde 60% (1.090) dos dados eram referentes a pacientes curados e 40% (727) a pacientes internados. A segunda base de dados (Base B')estava disposta em uma proporção de 70/30, sendo 70% (1.696) curados e 30% (727) internados.

A terceira base de dados (Base C') foi organizada em uma proporção de 50/50, onde 50% (1.090) dos dados foram divididos nas duas classes. Ao final, a quarta base de dados (Base D') também estava disposta em uma proporção de 50/50, sendo 50% (1.696) dos dados divididos nas duas classes. Na Tabela 4 são descritas as informações sobre as bases de dados a serem utilizadas na concepção do modelo preditivo do risco de internação.

Tabela 4 – Descrição das versões balanceadas das bases de dados a serem utilizadas na construção do modelo de predição da necessidade de internação.

Identificador	**Instâncias**	**Curado**	**Internado**
Base A'	1817 (100%)	1090 (60%)	727 (40%)
Base B'	2423 (100%)	1696 (70%)	727 (30%)
Base C'	2180 (100%)	1090 (50%)	1090 (50%)
Base D'	3392 (100%)	1696 (50%)	1696 (50%)

Fonte: Autoria própria.

4.6 Considerações Finais

Como apresentado, o conjunto de etapas necessárias para o pré-processamento e análise da base de dados se assemelha ao processo de Ciência de Dados, de modo a extrair conhecimento dos dados ao avaliar os diferentes tipos de informações, tais como: dados

pessoais, sintomas, comorbidades, histórico de vacinação e o estado de saúde dos pacientes. Mesmo com o desbalanceamento dos dados foi possível gerar um conjunto de bases de dados distintas, com intuito de avaliar o desempenho dos modelos. Além disso, foi possível realizar a análise dos atributos, visando identificar possíveis relações entre os dados. No Capítulo seguinte são apresentados o processo de escolha dos algoritmos e o desempenho alcançado nas diferentes bases de dados geradas.

5 Análise de Desempenho

Neste capítulo são apresentados quais foram os algoritmos selecionados para serem avaliados em relação ao desenvolvimento do modelo de predição (Seção 5.1). Em seguida, são apresentados os resultados alcançados na criação do modelo de predição da mortalidade de pacientes com a COVID-19 (Seção 5.2). Na sequência são ilustrados os resultados obtidos através do desenvolvimento do modelo de predição da probabilidade de internação de pacientes com a COVID-19 (Seção 5.3). Após isso, é apresentado o processo de disponibilização dos modelos (Seção 5.4. Por fim, são discutidos os resultados alcançados (Seção 5.5).

5.1 Algoritmos Selecionados

Tendo em vista a quantidade de algoritmos de aprendizado de máquina existentes, como forma melhorar o processo de seleção e avaliação dos algoritmos a serem utilizados no desenvolvimento dos modelos, foi empregada a biblioteca PyCaret (ALI, 2020). Essa biblioteca é baseada no aprendizado de máquina automatizado (também chamado de *AutoML*), que se resume no processo de automatizar as tarefas manuais executadas na construção dos modelos, a citar: o treinamento e avaliação dos algoritmos.

Para isso, o PyCaret dispõe de um conjunto de 14 algoritmos pré-selecionados, sendo: *Gradient Boosting*, *Light Gradient Boosting Machine*, *Logistic Regression*, *Linear Discriminant Analysis*, *Ridge*, *Ada Boost*, *Random Forest*, *SVM*, *K-NN*, *Extra Tree*, *Decision Tree*, *Naive Bayes*, *Quadratic Discriminant Analysis* e *Dummy*. Assim, por meio desse conjunto de algoritmos, é possível averiguar o desempenho de todos os algoritmos em relação a uma base de dados de entrada. Dessa forma, a partir da base de dados de entrada, uma saída é gerada contendo o ranking dos algoritmos com melhores resultados nas métricas de: *i)* Acurácia, *ii)* Precisão, *iii) Recall*, *iv) F1-Score* e v) AUC.

A partir desta definição, foi possível avaliar o desempenho do conjunto dos algoritmos da biblioteca PyCaret utilizando as bases de dados geradas na Seção 4.5 como entrada. Nos Apêndices B e C são apresentados os desempenhos dos dez melhores algoritmos em relação à cada uma das novas bases de dados de óbitos e internados, respectivamente. Assim, ao avaliar os 14 algoritmos de aprendizado de máquina supracitados, foi observado que os algoritmos *AdaBoost*, *Logistic Regression*, *Random Forest* e *Gradient Boosting* obtiveram os melhores resultados dentre os demais, tendo em vista o conjunto de métricas de desempenho.

No Quadros 6 e 7 são apresentados os algoritmos e seus parâmetros a serem

utilizados na construção dos modelos de predição do risco de mortalidade e internação, respectivamente. A partir dessa definição, os algoritmos selecionados serão avaliados a partir das respectivas bases de dados, com intuito de compreender os atributos estatisticamente mais relevantes e os resultados alcançados em relação a predição individual de cada uma das classes.

Para isso, conforme descrito na Seção 4.1, será utilizado a abordagem *hold-out*, que divide cada uma das bases de dados em duas partes, treinamento e teste. Neste trabalho foi utilizada a abordagem de 70% dos dados para treino e 30% para a realização dos testes. Os resultados obtidos pelos algoritmos são detalhados nas Seções a seguir.

5.2 Modelo de Agravamento 01: Predição da Mortalidade

Com a definição das bases de dados a serem utilizadas em cada um dos modelos, foram treinados, testados e avaliados cada um dos algoritmos escolhidos. Na Tabela 5 são apresentados os resultados obtidos pelos algoritmos utilizando as quatro versões da base de dados em relação às métricas de Acurácia "*A()*", Precisão "*P()*", *Recall* "*R()*" e *F1-score* "*F()*". Com intuito de averiguar os resultados de forma detalhada, foram destacados os resultados das métricas para ambas as classes, sendo 0 (Óbito) e 1 (Curado).

Ao analisar os resultados obtidos pelos algoritmos utilizando as duas primeiras versões das bases de dados (Bases A e B), é possível observar que o AB, o LR e o GB apresentaram desempenhos semelhantes em todas as métricas, a citar a acurácia média (82%) e a precisão média (83%). Em contrapartida, o RF apresentou desempenho inferior se comparado com os outros algoritmos, tendo em vista a acurácia média (79%) e a precisão média (77%). Outro fato a ser destacado está relacionado ao desempenho por classe, no qual, todos os algoritmos conseguiram um melhor resultado nas métricas de *Recall* e *F1-score* para predição de pacientes curados ($R(1)$ e $F(1)$), se comparado aos pacientes que vieram a óbito ($R(0)$ e $F(0)$).

Além disso, ao avaliar os resultados alcançados utilizando as duas outras versões da base de dados (Bases C e D), é possível constatar que o RF obteve o melhor desempenho em todas as métricas, a citar a acurácia média (81%) e a precisão média (83%). Enquanto isso, os algoritmos AB, LR e GB alcançaram um menor desempenho, mas com resultados semelhantes. Além disso, ao contrário do que foi observado nas duas primeiras bases de dados (Bases A e B), é válido destacar que os todos os algoritmos obtiveram resultados aproximados nas métricas de *Recall* e *F1-score* para classificação de pacientes curados e que vieram a óbito.

Posteriormente, observou-se o desempenho dos algoritmos em relação à métrica AUC *(Area Under Curve*), conforme apresentado na Figura 16. Ao analisar os resultados, foi constatado que assim como nas outras métricas, os algoritmos AB, LR e GB alcançaram

Quadro 6 – Algoritmos e seus respectivos parâmetros a serem avaliados no desenvolvimento do Modelo de Predição do risco de Mortalidade.

Algoritmo	**Parâmetros**
AdaBoost (AB)	algorithm = 'SAMME.R', base_estimator = None, learning_rate = 1.0, n_estimators = 50, random_state = 7846
Random Forest (RF)	bootstrap = True, ccp_alpha = 0.0, class_weight = None, criterion = 'gini', max_depth = None, max_features = 'auto', max_leaf_nodes = None, max_samples = None, min_impurity_decrease = 0.0, min_impurity_split = None, min_samples_leaf = 1, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 100, n_jobs = -1, oob_score = False, random_state = 7846, verbose = 0, warm_start = False
Logistic Regression (LR)	C = 1.0, class_weight = None, dual = False, fit_intercept = True, intercept_scaling = 1, l1_ratio = None, max_iter = 1000, multi_class = 'auto', n_jobs = None, penalty = 'l2', random_state=7846, solver = 'lbfgs', tol = 0.0001, verbose = 0, warm_start = False
Gradient Boosting (GB)	ccp_alpha = 0.0, criterion = 'friedman_mse', init = None, learning_rate = 0.1, loss = 'deviance', max_depth = 3, max_features = None, max_leaf_nodes = None, min_impurity_decrease = 0.0, min_samples_leaf = 1, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 100, n_iter_no_change = None, random_state = 1859, subsample = 1.0, tol = 0.0001, validation_fraction = 0.1, verbose = 0, warm_start = False

Fonte: Autoria própria.

resultados equivalentes em todas as cenários, obtendo um resultado médio de 0.86, 0.85, 0.84 e 0.85, nas bases de dados A, B, C e D, respectivamente. Em relação ao RF, o mesmo apresentou um desempenho abaixo dos outros algoritmos nas bases de dados A (0.83) e B (0.81). No entanto, o RF atingiu os melhores resultados nas bases de dados C (0.86) e D (0.89).

Quadro 7 – Algoritmos e seus respectivos parâmetros a serem avaliados no desenvolvimento do Modelo de Predição do risco de Internação.

Algoritmo	**Parâmetros**
AdaBoost (AB)	algorithm = 'SAMME.R', base_estimator = None, learning_rate = 1.0, n_estimators = 50, random_state = 2982
Random Forest (RF)	bootstrap = True, ccp_alpha = 0.0, class_weight = None, criterion = 'gini', max_depth = None, max_features = 'auto', max_leaf_nodes = None, max_samples = None, min_impurity_decrease = 0.0, min_impurity_split = None, min_samples_leaf = 1, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 100, n_jobs = -1, oob_score = False, random_state = 2982, verbose = 0, warm_start = False
Logistic Regression (LR)	C = 1.0, class_weight = None, dual = False, fit_intercept = True, intercept_scaling = 1, l1_ratio = None, max_iter = 1000, multi_class = 'auto', n_jobs = None, penalty = 'l2', random_state = 2982, solver = 'lbfgs', tol = 0.0001, verbose = 0, warm_start = False
Gradient Boosting (GB)	ccp_alpha = 0.0, criterion = 'friedman_mse', init = None, learning_rate = 0.1, loss = 'deviance', max_depth = 3, max_features = None, max_leaf_nodes = None, min_impurity_decrease = 0.0, min_samples_leaf = 1, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 100, n_iter_no_change = None, random_state = 2982, subsample = 1.0, tol = 0.0001, validation_fraction = 0.1, verbose = 0, warm_start = False

Fonte: Autoria própria.

Em seguida, foi utilizada a técnica SHAP por meio da biblioteca de código aberto de mesmo nome, com intuito de compreender a correlação entre o valor dos atributos e o resultado da predição, juntamente com o nível de importância de cada um dos atributos (LUNDBERG, 2022). A partir desta análise foi possível reduzir a quantidade de atributos das bases de dados, visando remover os atributos de menor poder preditivo no resultado.

Tabela 5 – Desempenho dos algoritmos na predição do risco de mortalidade dos pacientes com COVID-19.

Bases e Algoritmos	A	P (0)	R (0)	F (0)	P (1)	R (1)	F (1)
Base A							
AdaBoost	81,66	84,49	69,45	76,24	80,14	90,63	85,06
Random Forest	78,03	78,44	66,37	71,90	77,79	86,59	81,96
Logistic Regression	81,19	82,69	70,33	76,01	80,35	89,18	84,53
Gradient Boosting	81,47	83,86	69,67	76,11	80,17	90,15	84,87
Base B							
AdaBoost	83,86	83,38	60,49	70,12	84,00	94,51	88,94
Random Forest	81,13	76,18	57,81	65,74	82,68	91,76	86,98
Logistic Regression	83,51	81,93	60,71	69,74	83,99	93,90	88,66
Gradient Boosting	83,79	84,18	59,38	69,63	83,68	94,91	88,94
Base C							
AdaBoost	77,64	82,19	72,29	76,92	73,86	83,33	78,31
Random Forest	80,05	82,88	77,26	79,97	77,43	83,01	80,12
Logistic Regression	78,26	81,48	74,85	78,02	75,37	81,89	78,49
Gradient Boosting	78,42	83,05	73,04	77,72	74,57	84,13	79,07
Base D							
AdaBoost	78,94	81,99	74,50	78,07	76,38	83,43	79,75
Random Forest	83,38	83,99	82,74	83,36	82,79	84,04	83,41
Logistic Regression	78,54	81,01	74,90	77,84	76,40	82,23	79,21
Gradient Boosting	79,79	82,18	76,39	79,18	77,69	83,23	80,37

Legenda: *Acurácia A(), Precisão P(), Recall R(), F1-Score F(), Óbito (0) e Curado (1).*

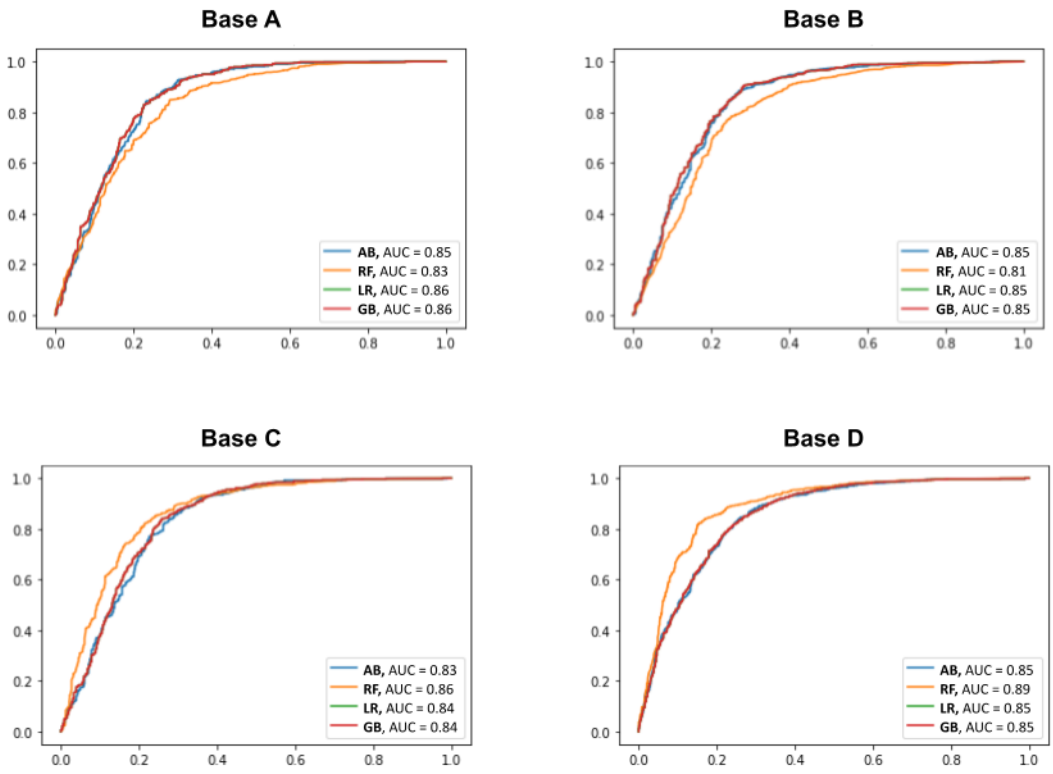
Fonte: Autoria própria.

Na Figura 17 são apresentados os resultados da correlação dos principais atributos em relação a classificação da evolução do estado clínico do paciente (óbito ou curado). Por meio da análise gráfica, um atributo foi considerado de alta correlação quando observou-se uma maior dispersão entre os pontos. Como caracterizado na legenda, cada um dos pontos no gráfico representa os atributos de um paciente, podendo assumir valores em uma escala de azul (valor mais baixo) à vermelho (valor mais alto).

A escala de valores pode variar em cada um dos atributos, por exemplo, o atributo *dorDeCabeca* possui apenas dois valores: *i)* 0, representado em azul por ser o valor mais baixo, que indica que o paciente não estava com dor de cabeça e *ii)* 1, representado em vermelho por ser o valor mais alto, assinalando que o paciente apresentou dor de cabeça. Já o atributo *qntVacinas* está representado em três valores: *i)* 0, o menor valor (cor azul), indicando que o paciente não está vacinado; *ii)* 1, o valor mediano (cor intermediária entre azul e vermelho), destacando que o paciente foi vacinado com uma dose e *iii)* 2, o valor mais alto (cor vermelha), ressaltando que o paciente recebeu duas doses da vacina.

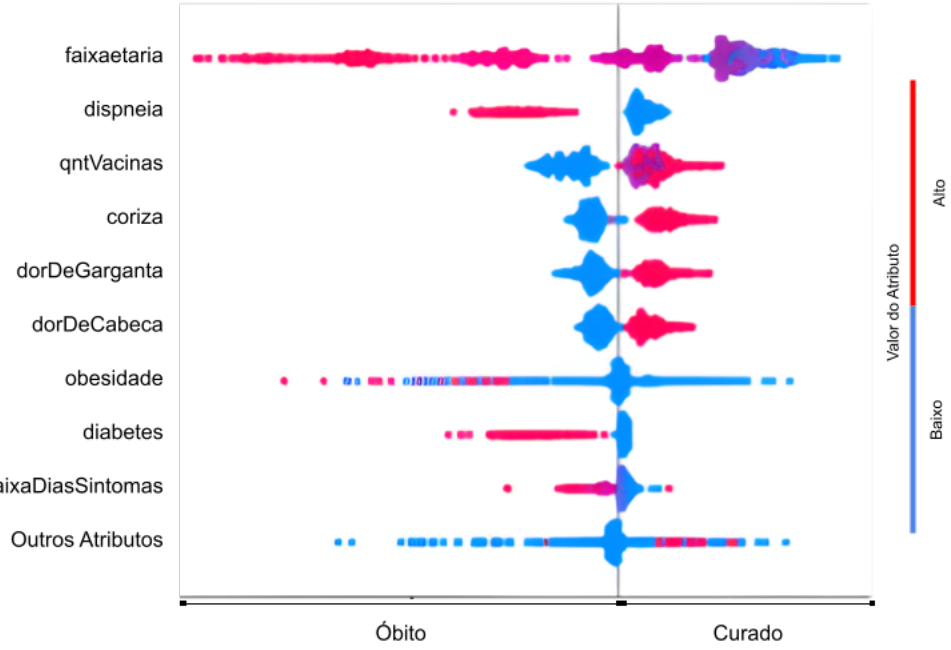
O atributo *faixaEtaria* possui um conjunto de valores de 0 até 7, sendo 0 o valor

Figura 16 – Desempenho dos algoritmos em relação à métrica AUC utilizando as Bases A, B, C e D.



Fonte: Autoria própria.

Figura 17 – Análise SHAP da correlação dos atributos em relação ao risco de mortalidade.



Fonte: Autoria própria.

mais baixo e 7 o valor mais alto. Nesse contexto, é possível observar na Figura 17 que pacientes com uma idade elevada (cor vermelha) possuem uma maior probabilidade de óbito, enquanto pacientes mais jovens (cor azul) tendem a ser classificados como curados. Em relação ao atributo *dispneia*, nota-se que os pacientes que vieram a óbito tendem a apresentar este sintoma (cor vermelha). No cenário de vacinação, a partir do atributo *qntVacinas* é observado que os pacientes não vacinados possuem uma maior probabilidade de mortalidade, enquanto os pacientes que foram vacinados com uma ou duas doses possuem uma maior chance de serem curados.

Ao observar os atributos *coriza*, *dorDeGarganta* e *dorDeCabeca*, pode-se verificar uma similaridade no comportamento dos valores. No qual, pacientes que vieram a óbito possuem uma probabilidade maior de não apresentar os sintomas (cor azul), enquanto os pacientes curados tendem a apresentar os sintomas (cor vermelha). Por fim, em relação aos atributos *diabetes* e *faixaDiasSintomas*, nota-se que os pacientes diabéticos e que apresentam os sintomas característicos da COVID-19 há mais dias possuem uma maior probabilidade de mortalidade.

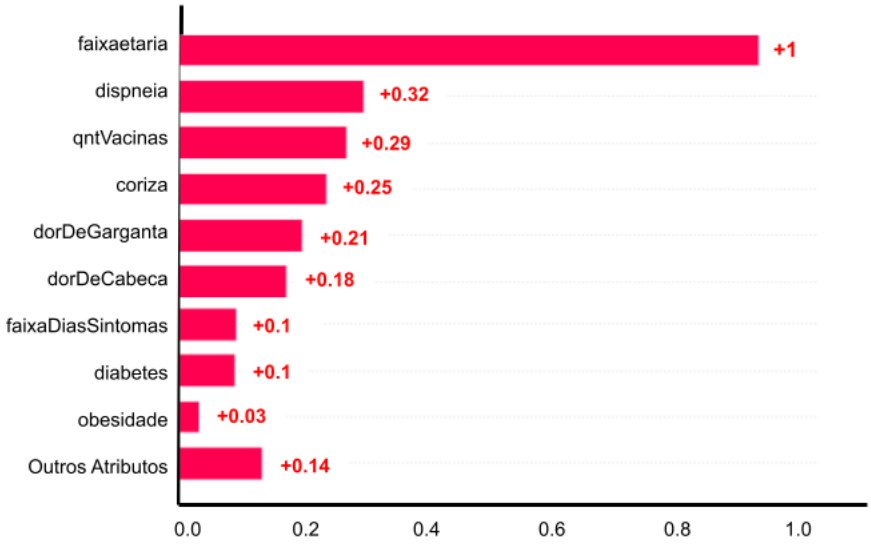
Posteriormente, foi calculado o nível de importância dos atributos em relação a evolução do estado de saúde dos pacientes, como ilustrado na Figura 18. O nível de importância é medido em uma escala de 0 à 1, sendo 0 o menor nível e 1 o maior nível. Mediante os resultados, é possível observar uma similaridade entre os atributos de maior correlação destacados anteriormente e os que possuem um maior nível de importância. Dentre os atributos de maior nível de importância, destacam-se: i) os atributos demográficos: *faixaEtaria* e *qntVacinas*; ii) os sintomas: *dispneia*, *dorDeGarganta*, *dorDeCabeca* e *dorNoCorpo*; e iii) as comorbidades: *obesidade* e *diabetes*.

Em seguida, como forma de avaliar o desempenho dos algoritmos utilizando apenas os atributos de maior importância identificados, foi realizada a redução de atributos. Para isso, foram selecionados os atributos com nível de importância maior ou igual a 0.1, sendo escolhidos: *faixaEtaria*, *dispneia*, *qntVacinas*, *coriza*, *dorDeGarganta*, *dorDeCabeca*, *faixaDiasSintomas* e *diabetes*, além do atributo alvo *evolucaoCaso*.

Com isso, os algoritmos foram novamente avaliados em relação às métricas de desempenho, conforme apresentado na Tabela 6. A partir da análise de desempenho, notou-se que os resultados obtidos não divergiram em relação a utilização das bases de dados contendo todos os atributos (Tabela 5). Ao avaliar o desempenho em relação a primeira base de dados (Base A), observa-se que o RF alcançou um desempenho inferior se comparado com o AB, LR e GB, que obtiveram resultados semelhantes em todas as métricas, a citar a acurácia média de 81% e precisão média de 79%. Entretanto, nas três bases de dados seguintes, todos os algoritmos apresentam resultados similares em todas as métricas, destacando-se a acurácia média de 79% e precisão média de 80%.

Também é válido destacar que mesmo com a redução dos atributos, o resultado

Figura 18 – Análise SHAP do importância dos atributos em relação ao risco de mortalidade.



Fonte: Autoria própria.

das métricas de *Recall* e *F1-score* nas bases de dados A e B não se manteve equilibrado na predição das duas classes, destacando-se o *Recall* médio de 65% e *F1-score* médio de 72% na predição de pacientes que vieram a óbito, e o *Recall* médio de 91% e *F1-score* médio de 86% na predição de pacientes curados. No entanto, nas duas últimas bases de dados (C e D), foi constatado um maior equilíbrio nas métricas, onde foi possível observar um *Recall* médio de 74% e 80% e *F1-score* de 77% e 78%, na predição de pacientes que vierem a óbito e foram curados, respectivamente.

Posteriormente, foram calculados os valores obtidos em relação à métrica AUC, conforme ilustrado na Figura 19. Mediante a análise dos valores, é possível constatar que novamente os algoritmos AB, LR e GB alcançaram os melhores resultados nas bases de dados A (0.86) e B (0.85). De forma similar às outras métricas, o RF obteve um valor abaixo (0.82) dos outros três algoritmos nas bases de dados A e B. No entanto, nas duas últimas bases de dados todos o algoritmos obtiveram desempenho equivalentes, sendo 0.84 e 0.85 nas bases de dados C e D, respectivamente.

Mediante a avaliação dos algoritmos a partir do conjunto de métricas de desempenho, foi possível observar que os resultados obtidos foram estatisticamente similares. No entanto, foi possível identificar que o algoritmo GB, treinado a partir da Base C, alcançou um bom desempenho e resultados aproximados nas métricas de Precisão, *Recall* e *F1-score* para ambas as classes. Diante disso, o modelo treinado a partir do algoritmo GB foi selecionado como o modelo de predição da probabilidade de mortalidade dos pacientes com COVID-19.

Assim, com intuito de validar o desempenho deste modelo de predição em relação a outros dados, foram selecionadas as cinco bases de dados dos estados do Brasil com

Tabela 6 – Desempenho dos algoritmos na predição do risco de mortalidade utilizando a redução dos atributos.

Bases e Algoritmos	A	P (0)	R (0)	F (0)	P (1)	R (1)	F (1)
Base A							
AdaBoost	81,38	84,55	68,57	75,73	79,72	90,79	84,89
Random Forest	78,49	79,02	67,03	72,53	78,20	86,91	82,33
Logistic Regression	81,01	82,43	70,11	75,77	80,20	89,01	84,38
Gradient Boosting	81,10	83,87	68,57	75,45	79,63	90,31	84,63
Base B							
AdaBoost	83,37	83,02	58,93	68,93	83,47	94,51	88,65
Random Forest	82,32	78,59	59,82	67,93	83,49	92,57	87,80
Logistic Regression	83,44	81,49	60,94	69,73	84,03	93,69	88,60
Gradient Boosting	83,86	83,38	60,49	70,12	84,00	94,51	88,94
Base C							
AdaBoost	77,64	81,33	73,49	77,22	74,42	82,05	78,05
Random Forest	75,62	78,50	72,59	75,43	73,00	78,85	75,81
Logistic Regression	77,80	81,40	73,80	77,41	74,64	82,05	78,17
Gradient Boosting	77,95	82,20	73,04	77,35	74,36	83,17	78,52
Base D							
AdaBoost	77,94	80,50	74,11	77,17	75,74	81,83	78,67
Random Forest	77,69	79,37	75,20	77,23	76,17	80,22	78,14
Logistic Regression	77,30	78,83	75,00	76,87	75,89	79,62	77,71
Gradient Boosting	78,89	81,42	75,20	78,18	76,70	82,63	79,56

** Sendo Acurácia A(), Precisão P(), Recall R() e F1-Score F(). E as classificações de Óbito (0) e Curado (1).*

Fonte: Autoria própria.

maior quantitativo de instâncias (Tabela 7). Em seguida, diante do desempenho do modelo (Tabela 8), notou-se que os resultados foram semelhantes aos obtidos durante o treinamento, fato esse que enfatiza a estabilidade, qualidade e generalidade do modelo, para outros conjuntos de dados.

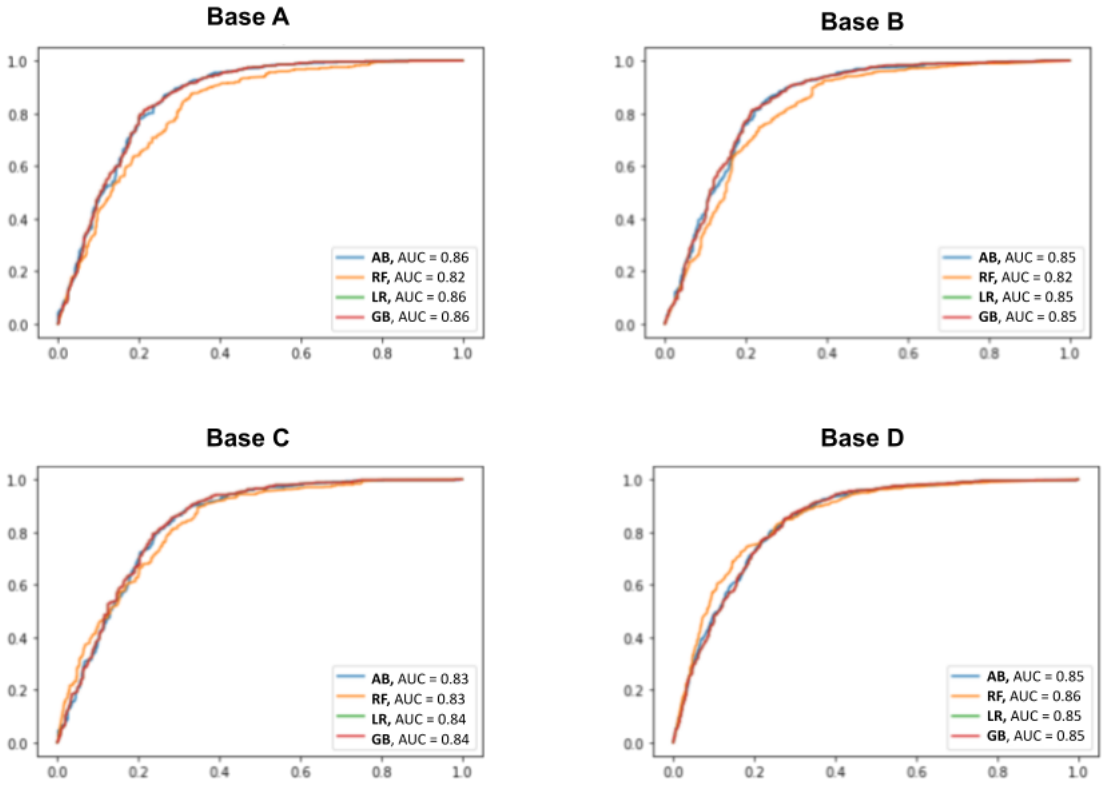
Tabela 7 – Descrição das cinco bases com maior quantitativo de instâncias usadas na etapa de validação do modelo.

Base UF	**Instâncias**	**Óbito (0)**	**Curado (1)**
MG	4863	1215	3648
BA	3847	962	2885
RJ	1928	482	1446
SC	1580	395	1185
RS	876	219	657

Fonte: Autoria própria.

Em relação às métricas escolhidas para a avaliação dos modelos, foi observado que o modelo de predição da mortalidade alcançou valores superiores a 80% nas médias da Acurácia e AUC. Ao avaliar a métrica de Precisão, é possível verificar que o modelo obteve

Figura 19 – Desempenho dos algoritmos em relação à métrica AUC utilizando as bases A, B, C e D



Fonte: Autoria própria.

um resultado melhor na classificação de pacientes curados (88%), se comparado com os pacientes que vieram a óbito (76%).

Tabela 8 – Resultados da Validação do Modelo de Predição da Mortalidade em relação às bases de dados dos outros estados do Brasil.

Base (UF)	A	AUC	P (0)	R (0)	F (0)	P (1)	R (1)	F (1)
MG	85,61	0,88	75,94	76,13	76,04	89,76	89,67	89,72
BA	88,12	0,93	77,96	84,20	80,96	92,99	89,80	91,36
RJ	84,49	0,88	73,63	75,31	74,46	89,30	88,42	88,86
SC	84,80	0,88	74,68	74,68	74,68	89,14	89,14	89,14
RS	81,37	0,80	78,62	52,05	62,64	82,05	93,93	87,59
Média	84,88	0,87	76,17	72,47	73,76	88,65	90,19	89,33

Legenda: *Acurácia A(), Precisão P(), Recall R(), F1-Score F(), Óbito (0) e Curado (1).*

Fonte: Autoria própria.

Também é válido destacar os resultados alcançados para *Recall* e *F1-score* em relação aos pacientes curados (aproximadamente 90% e 89%, respectivamente) e que vieram a óbito (aproximadamente 72% e 74%, respectivamente). A equidade nos resultados das

métricas para ambas as classes é de fundamental importância, pois é possível constatar que o modelo conseguiu classificar de forma similar as instâncias, mesmo utilizando dados distintos. Com o término da etapa de validação, foi criada uma aplicação Web (Seção 5.4) para demonstrar a aplicabilidade prática do modelo de predição do risco de mortalidade.

5.3 Modelo de Agravamento 02: Predição da Internação

Assim como o modelo de predição do risco de mortalidade, na Tabela 9 são apresentados os resultados obtidos pelos algoritmos utilizando as quatro versões da base de dados em relação às métricas de Acurácia "*A()*", Precisão "*P()*", *Recall* "*R()´´* e *F1-score* "*F()*". Com intuito de averiguar os resultados de forma detalhada, foram destacados os resultados das métricas para ambas as classes, sendo 0 (Internado) e 1 (Curado).

Ao analisar os resultados alcançados utilizando a Base A', é possível observar que o GB obteve o melhor desempenho, a citar a acurácia média de 74% e precisão média de 74%. Ao avaliar os resultados alcançados utilizando as duas últimas versões da base de dados Bases (C' e D'), notou-se que o algoritmo RF obteve o melhor desempenho em todas as métricas, obtendo uma acurácia e precisão média de 76% e 77%, respectivamente. Além disso, foi observado um equilíbrio na classificação de ambas as classes, tendo em vista o *Recall* médio de 78% e 75% e *F1-score* de 77% e 75%, na predição de pacientes internados e curados, respectivamente.

Em seguida, foi destacado o desempenho dos algoritmos em relação à métrica AUC, como ilustrado na Figura 20. Ao analisar os resultados, foi constatado uma similaridade com o modelo de predição da mortalidade (Seção 5.2), no qual, os algoritmos AB, LR e GB alcançaram o melhores resultados nas bases A' (0.76) e B' (0.72). No entanto, o RF atingiu os melhores resultados nas bases de dados C' (0.82) e D' (0.85).

Assim como o modelo anterior, foi utilizada a técnica SHAP, com intuito de compreender a correlação entre o valor dos atributos e o resultado da predição. Na Figura 21 são apresentados os resultados da correlação dos principais atributos em relação a classificação da evolução do estado clínico do paciente (internado ou curado). Mediante a análise, observou-se a partir do atributo *faixaEtaria* que pacientes com uma idade elevada (cor vermelha) possuem uma maior probabilidade de internação. Em relação ao atributo *dispneia*, nota-se os pacientes que foram internados tendem a apresentar o sintoma (cor vermelha).

No cenário de vacinação, a partir do atributo *qntVacinas* é observado que os pacientes não vacinados possuem uma maior probabilidade de internação. Ao avaliar os atributos *coriza*, *dorDeGarganta* e *dorDeCabeca*, foi constatado uma similaridade no comportamento dos valores, no qual, pacientes que foram internados tendem a não apresentar os sintomas (cor azul), enquanto os pacientes curados tendem a apresentar

Tabela 9 – Desempenho dos algoritmos na predição do risco de internação dos pacientes com COVID-19.

Bases e Algoritmos	A	P (0)	R (0)	F (0)	P (1)	R (1)	F (1)
Base A'							
AdaBoost	71,61	67,42	55,30	60,76	73,64	82,37	77,76
Random Forest	70,51	64,89	56,22	60,25	73,46	79,94	76,56
Logistic Regression	71,43	69,18	50,69	58,51	72,35	85,11	78,21
Gradient Boosting	74,73	71,82	59,91	65,33	76,16	84,50	80,12
Base B'							
AdaBoost	74,83	67,33	42,98	52,47	76,78	90,04	82,88
Random Forest	74,69	65,45	45,96	54,00	77,40	88,41	82,54
Logistic Regression	75,24	71,32	39,15	50,55	76,09	92,48	83,49
Gradient Boosting	75,38	70,29	41,28	52,01	76,57	91,67	83,44
Base C'							
AdaBoost	72,94	76,49	68,55	72,30	69,89	77,60	73,54
Random Forest	75,99	77,44	75,37	76,39	74,54	76,66	75,58
Logistic Regression	72,78	75,56	69,73	72,53	70,26	76,03	73,03
Gradient Boosting	74,92	76,95	73,29	75,08	72,97	76,66	74,77
Base D'							
AdaBoost	72,30	75,05	69,94	72,41	69,71	74,85	72,19
Random Forest	77,11	76,62	80,53	78,53	77,71	73,42	75,50
Logistic Regression	70,43	71,59	71,46	71,52	69,18	69,33	69,25
Gradient Boosting	73,67	74,95	74,10	74,52	72,32	73,21	72,76

Legenda: *Acurácia A(), Precisão P(), Recall R(), F1-Score F(), Internado (0) e Curado (1).*

Fonte: Autoria própria.

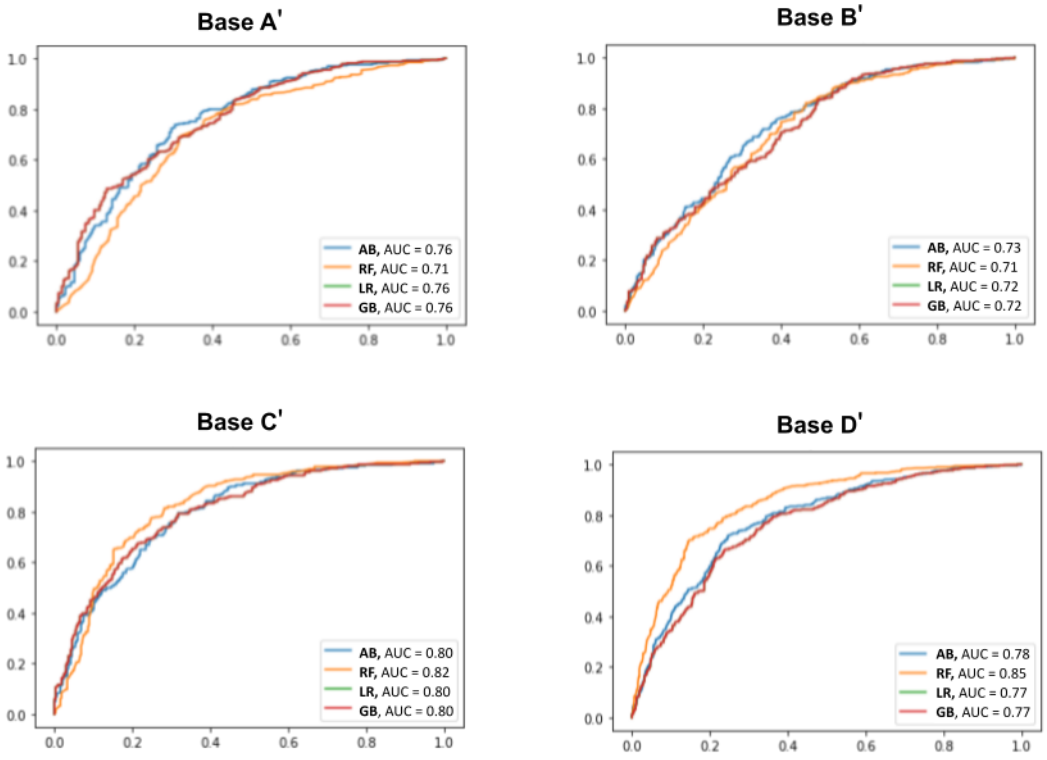
os sintomas (cor vermelha). Por fim, nota-se que os pacientes que apresentaram *febre* possuem uma maior probabilidade de internação.

Posteriormente, foi calculado o nível de importância dos atributos em relação a evolução do estado de saúde dos pacientes (Figura 22). Mediante os resultados, foi possível observar uma similaridade entre os atributos de maior correlação e os que possuem um maior nível de importância. Dentre os atributos de maior nível de importância, destacam-se: i) os atributos demográficos: *faixaEtaria* e *qntVacinas*; ii) os sintomas: *dorDeCabeca*, *dorDeGarganta*, *dispneia*, *tosse*, *faixaDiasSintomas* e *coriza*; e iii) as comorbidades: *renal* e *diabetes*.

Em seguida, como forma de avaliar o desempenho dos algoritmos utilizando apenas atributos de maior importância, foi realizada a redução do atributos. Para isso, foram selecionados os atributos com nível de importância maior ou igual a 0.1, resultando em um total de 11 atributos: *faixaEtaria*, *dorDeCabeca*, *dorDeGarganta*, *renal*, *dispneia*, *diabetes*, *tosse*, *qntVacinas*, *faixaDiasSintomas* e *coriza*, além do atributo alvo *evolucaoCaso*.

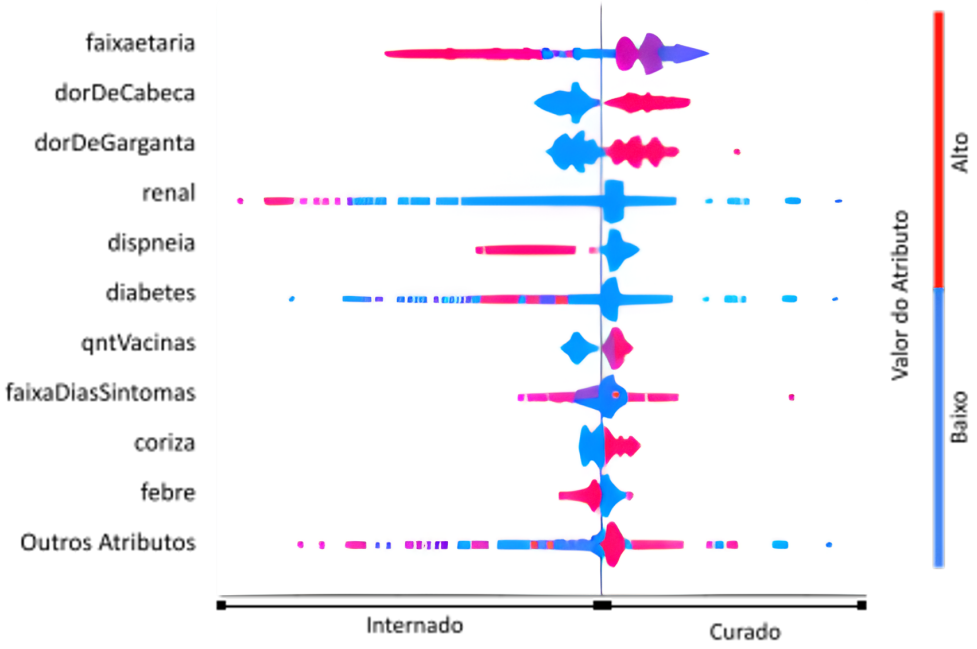
Com isso, os algoritmos foram novamente avaliados em relação às métricas de

Figura 20 – Desempenho dos algoritmos em relação à métrica AUC utilizando as Base A', B', C' e D'



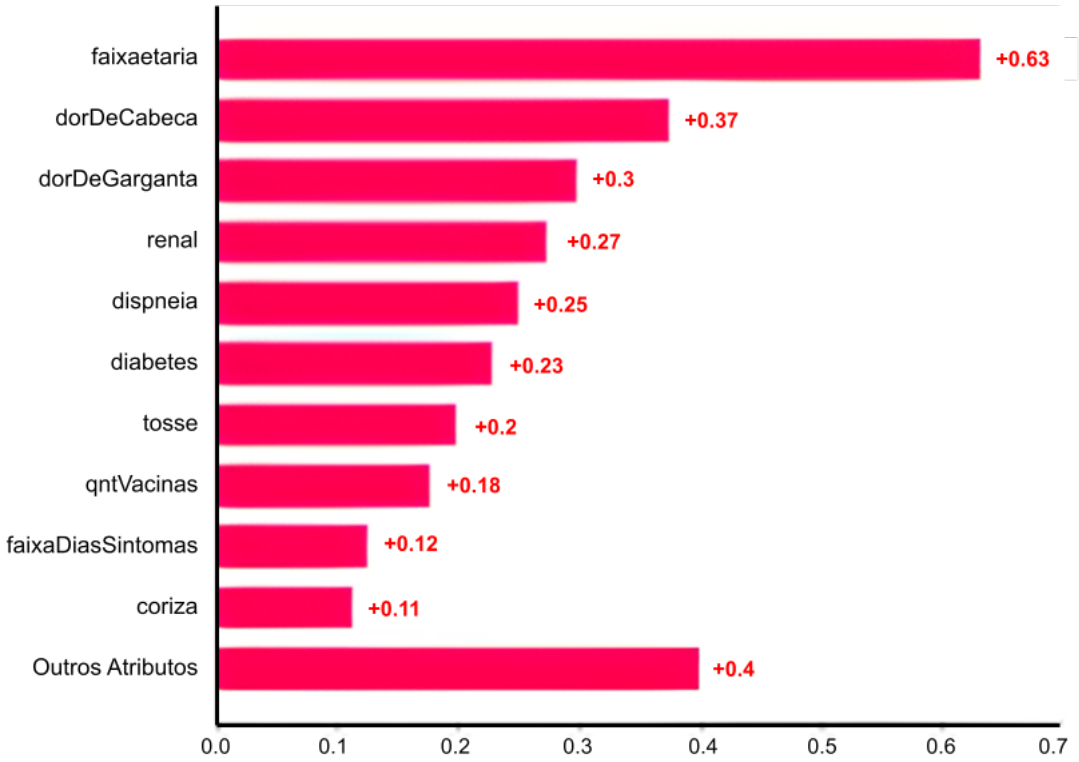
Fonte: Autoria própria.

Figura 21 – Análise SHAP da correlação dos atributos em relação ao risco de internação.



Fonte: Autoria própria.

Figura 22 – Análise SHAP do nível de importância dos atributos em relação ao risco de internação.



Fonte: Autoria própria.

desempenho, conforme apresentado na Tabela 6. Ao avaliar o desempenho dos algoritmos em relação a Base A', observou-se que os algoritmos AB e GB alcançaram os melhores resultados, obtendo uma a'curácia e precisão média de 71% e 70%, respectivamente. Em relação às bases de dados balanceadas utilizando a técnica SMOTE (Bases C' e D'), foi observado que o algoritmo GB obteve resultados superiores aos demais algoritmos em praticamente todas as métricas, destacando-se a acurácia média de 74% e a precisão média de 73%.

Posteriormente, foram calculados os valores obtidos em relação à métrica AUC (Figura 23). Em relação aos valores, foi possível constatar que os algoritmos AB, LR e GB alcançaram os melhores resultados na base de dados A (0.75). Já na base de dados B, o algoritmo AB obteve o melhor desempenho (0.73). Por fim, na base de dados C, todos os algoritmos alcançaram resultados similares (0.78), enquanto que na base de dados D o RF alcançou o maior valor (0.81).

Mediante a avaliação dos algoritmos a partir do conjunto das métricas de desempenho, foi possível observar que os resultados obtidos foram estatisticamente equilibrados. No entanto, foi possível identificar que o modelo de predição treinado a partir do algoritmo GB e da Base B', alcançou o melhor desempenho dentre os demais, sendo este selecionado como o modelo de predição da probabilidade de internação dos pacientes com COVID-19.

Tabela 10 – Desempenho dos algoritmos na predição do risco de internação utilizando a redução dos atributos.

Bases e Algoritmos	A	P (0)	R (0)	F (0)	P (1)	R (1)	F (1)
Base A'							
AdaBoost	72,16	69,01	54,38	60,82	73,60	83,89	78,41
Random Forest	66,48	58,76	52,53	55,47	70,74	75,68	73,13
Logistic Regression	68,68	64,20	47,93	54,88	70,57	82,37	76,02
Gradient Boosting	71,25	66,85	54,84	60,25	73,37	82,07	77,47
Base B'							
AdaBoost	74,42	68,15	39,15	49,73	75,84	91,26	82,84
Random Forest	74,14	62,98	48,51	54,81	77,84	86,38	81,89
Logistic Regression	73,59	67,48	35,32	46,37	74,83	91,87	82,48
Gradient Boosting	74,83	67,81	42,13	51,97	76,59	90,45	82,95
Base C'							
AdaBoost	72,17	75,92	67,36	71,38	69,01	77,29	72,92
Random Forest	73,70	75,86	71,81	73,78	71,64	75,71	73,62
Logistic Regression	71,71	74,84	67,95	71,23	68,97	75,71	72,18
Gradient Boosting	75,54	78,46	72,40	75,31	72,89	78,86	75,76
Base D'							
AdaBoost	71,61	74,49	69,00	71,64	68,94	74,44	71,58
Random Forest	72,69	73,46	74,29	73,87	71,84	70,96	71,40
Logistic Regression	70,43	72,35	69,75	71,03	68,50	71,17	69,81
Gradient Boosting	73,28	74,66	73,53	74,10	71,83	73,01	72,41

Legenda: *Acurácia A(), Precisão P(), Recall R(), F1-Score F(), Internado (0) e Curado (1).*

Fonte: Autoria própria.

Com intuito de validar o desempenho do modelo de predição em relação a outros dados, foram selecionadas as cinco bases de dados dos estados do Brasil com o maior quantitativo de instâncias (Tabela 11). Em seguida, a partir do desempenho do modelo (Tabela 12), notou-se que os resultados, de forma geral, foram semelhantes aos obtidos durante o treinamento, fato esse que enfatiza a estabilidade, qualidade e generalidade do modelo, para outros conjuntos de dados.

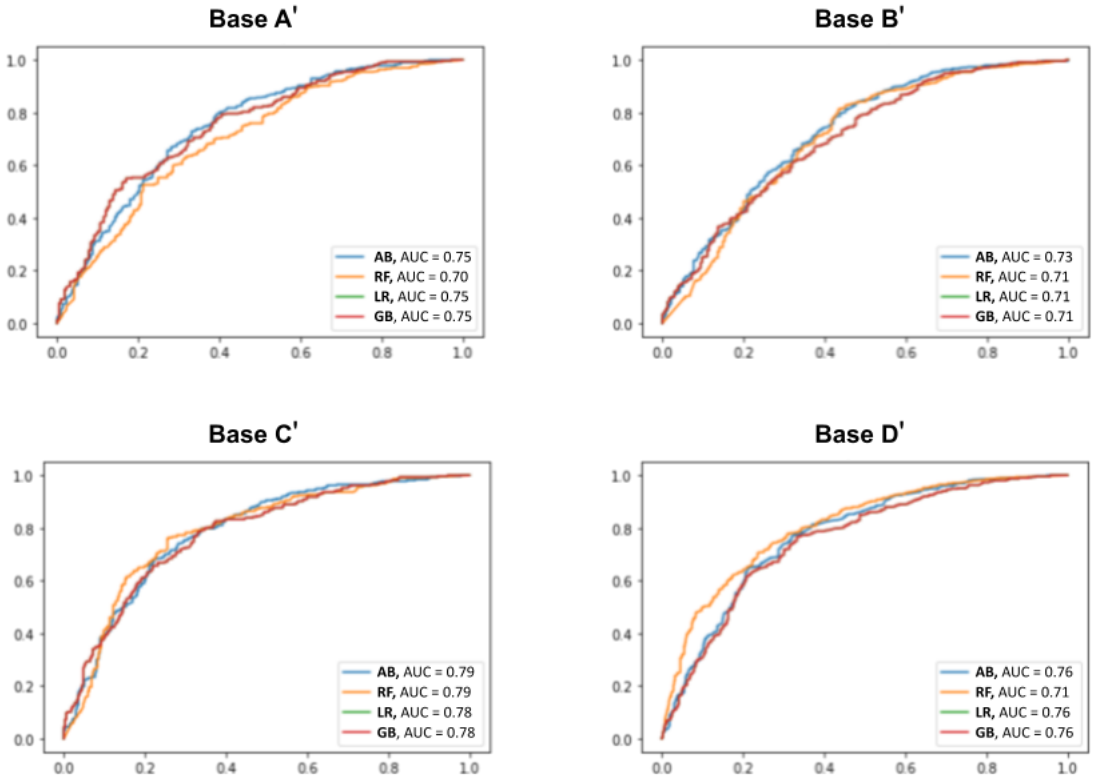
Tabela 11 – Descrição das cinco bases com maior quantitativo de instâncias usadas na etapa de validação do modelo.

Base UF	**Instâncias**	**Internado (0)**	**Curado (1)**
BA	1439	360	1079
MG	805	201	604
SC	804	201	603
RS	756	189	567
RJ	652	163	489

Fonte: Autoria própria.

Em relação às métricas, foi observado que o modelo alcançou valores superiores

Figura 23 – Desempenho dos algoritmos em relação à métrica AUC utilizando as Base A', B', C' e D'.



Fonte: Autoria própria.

a 75% nas médias da Acurácia e AUC (aproximadamente 81% e 78%, respectivamente). Ao avaliar a métrica de Precisão, é possível verificar que o modelo obteve um resultado melhor na classificação de pacientes curados (84%), se comparado com os pacientes que necessitaram de internação (67%).

Também é válido destacar os resultados alcançados para *Recall* e *F1-score* em relação aos pacientes curados (sendo aproximadamente 92% e 88%, respectivamente). Contudo, o modelo obteve resultados inferiores no *Recall* e *F1-score* em relação aos pacientes internados (aproximadamente 49% e 56%, respectivamente). Ao final, assim como o modelo de predição da mortalidade de pacientes com a COVID-19, foi criada uma aplicação Web (Seção 5.4) para demonstrar a aplicabilidade prática de uso do modelo de predição do risco de internação.

5.4 Disponibilização dos Modelos de Predição

Após o desenvolvimento e validação dos modelos de predição da probabilidade de agravamento em relação a internação e mortalidade dos pacientes, foi realizado o processo de disponibilização. Como forma de possibilitar o acesso por um maior número de usuários,

Tabela 12 – Resultados da Validação do Modelo de Predição da Risco de Internação em relação às bases de dados dos outros estados do Brasil.

Base (UF)	A	AUC	P (0)	R (0)	F (0)	P (1)	R (1)	F (1)
MG	82,34	0,78	68,32	54,73	60,77	85,85	91,54	88,6
BA	83,32	0,85	70,13	58,06	63,53	86,77	91,75	89,19
RJ	81,44	0,78	68,10	48,47	56,63	84,33	92,43	88,2
SC	79,85	0,77	63,45	45,77	53,18	83,46	91,21	87,16
RS	78,84	0,72	63,55	35,98	45,95	81,36	93,12	86,84
Média	81,16	0,78	66,71	48,60	56,01	84,35	92,01	88,00

Legenda: *Acurácia A(), Precisão P(), Recall R(), F1-Score F(), Internado (0) e Curado (1).*

Fonte: Autoria própria.

os modelos foram disponibilizados por meio de uma aplicação *Web*, permitindo o acesso via navegador em dispositivos móveis ou *desktop*.

Para realizar a disponibilização foi escolhido o pacote *Streamlit*, que possibilita a construção de aplicações Web a partir de código Python (STREAMLIT, 2022). O pacote em questão foi escolhido por ser um dos principais pacotes gratuitos disponíveis e por ter sido desenvolvido na mesma linguagem utilizada na análise dos dados e criação dos modelos de predição.

Para isso, inicialmente, os modelos de predição foram convertidos em serviços através do método *dump* do pacote *Joblib* (JOBLIB, 2022). Em seguida, foi desenvolvido um formulário para que os usuários possam informar os dados de entrada necessários para a predição do modelo, tais como: dados demográficos, sintomas, comorbidades e histórico de vacinação. Os dados necessários para a predição variam entre os dois modelos, como mostrado na Seções 5.2 e 5.3.

Assim, após o usuário informar os dados de entrada, será possível solicitar a predição do agravamento do estado clínico. O valor da probabilidade de agravamento nos dois modelos preditivos é calculado a partir do método *predict_proba* do pacote *Sklearn*, mesmo pacote utilizado no desenvolvimento do modelo de predição (PEDREGOSA *et al.*, 2011). A partir desse método, é feito o cálculo da probabilidade com base no número de amostras utilizadas no treinamento do algoritmo. Ao final dessas etapas, os dois modelos foram disponibilizados[1] [2].

Também é valido destacar que, como observado na avaliação dos algoritmos selecionados em relação às métricas de desempenho, foi constatado uma similaridade nos resultados alcançados. Diante disso, por padrão, o algoritmo escolhido na realização da predição do modelo foi o que alcançou os melhores resultados (em ambos os modelos, o

[1] Modelo de Predição da Mortalidade: https://cutt.ly/5Myl2Rn
[2] Modelo de Predição da Internação: https://cutt.ly/VMylJef

Gradient Boosting). No entanto, caso o usuário prefira utilizar outro algoritmo na realização da predição, é possível escolher o algoritmo logo antes de preencher os dados de entrada.

5.5 Discussões dos Resultados

Ao avaliar os atributos estatisticamente de maior importância nos modelos em relação a probabilidade de agravamento, foram encontradas algumas similaridades, como por exemplo: a idade avançada, dispneia, ausência de vacinação, diabetes e febre, em pacientes que foram internados ou vieram a óbito. Além das características de: dor de cabeça, dor de garganta e coriza, em pacientes que tiveram a doença mas foram curados.

Diante dessas informações, foi possível identificar trabalhos que alcançaram resultados similares em relação aos fatores de agravamento. Nos trabalhos de Kupeli e Yüksel (2020) e Singhal *et al.* (2021) foi observado que os pacientes diagnosticados com a COVID-19 que possuem mais de 60 anos são mais propensos a desenvolverem um quadro grave da doença e necessitar de ventilação mecânica. Já em Shi *et al.* (2020) foi identificado a partir da revisão dos resultados de 11 artigos que a dispneia estava significativamente associada a uma maior probabilidade de mortalidade em pacientes com a COVID-19. Além disso, no trabalho de Gul, Htun e Inayat (2020) os autores destacaram que mesmo a febre sendo um sintoma comum em diversas doenças, no contexto da COVID-19, os pacientes que apresentaram febre tendem a possuir uma maior probabilidade de necessidade de internação.

Ainda em relação aos fatores de agravamento, Baj *et al.* (2020) e Zayet *et al.* (2020) identificaram que pacientes que apresentaram dor de cabeça e dor de garganta possuem uma probabilidade maior de cura. Os autores enfatizaram que esse fato pode ser justificado pois os respectivos sintomas geralmente são comuns em outras doenças e possuem um tratamento mais simplificado, se comparado com outros sintomas da COVID-19. Ao final, como forma de qualificar os resultados obtidos através do modelo de predição da probabilidade de mortalidade de pacientes com a COVID-19, foi conduzida uma comparação com alguns dos trabalhos selecionados no Mapeamento Sistemático da Literatura (Capítulo 3), como mostrado na Tabela 13.

Ao comparar o trabalho de Booth, Abels e McCaffrey (2020), é possível averiguar que o modelo desenvolvido alcançou uma AUC maior que a do presente trabalho (0.93 e 0.89, respectivamente), no entanto, os autores utilizaram dados referentes à exames clínicos, o que limita a aplicabilidade do modelo. Ao analisar o trabalho de Das, Mishra e Gopalan (2020), constatou-se que o modelo obteve uma AUC menor se comparado com o presente trabalho. Além disso, os autores utilizaram apenas dados demográficos, o que impossibilita a investigação de informações fundamentais, tais como os sintomas e comorbidades dos pacientes.

Tabela 13 – Descrição das características dos trabalhos relacionados em relação ao presente trabalho.

Trabalho	**Dados**	**Algoritmo**	**Disponível**	**AUC**
Booth, Abels e McCaffrrey (2020)	Exames Clínicos	*Support Vector Machines*	Não	0.93
Das, Mishra e Gopalan (2020)	Demográficos	*Logistic Regression*	Sim	0.83
Kim *et al.* (2020)	Demográficos, Sintomas e Comorbidades	XGBoost	Sim	0.88
Modelo de Predição da Mortalidade	Demográficos, Histórico de Vacinação Sintomas e Comorbidades	Gradient Boosting	Sim	0.89

Fonte: Autoria própria.

Em relação ao trabalho de Kim *et al.* (2020), foi observado que o modelo obteve uma AUC similar a obtida neste trabalho (0.88 e 0.89, respectivamente). No entanto, além de utilizar dados demográficos, sintomas e comorbidades, o presente trabalho também avalia o histórico de vacinação dos pacientes. Ao final, um fato que vale a pena ser destacado é que em nenhum dos trabalhos relacionados foram descritos os resultados das métricas por classe, apenas o desempenho médio. Este fato é de fundamental importância, pois o modelo pode estar classificando bem uma classe e outra não, descaracterizando os resultados obtidos.

6 Conclusões

A pandemia do novo coronavírus impactou significativamente no aumento dos índices de internação e mortalidade em todo o mundo, tendo em vista a disseminação da doença e a dificuldade em prever os pacientes com maior risco de agravamento do quadro clínico.

Em meio a essa dificuldade, a partir da realização deste trabalho foi possível: *i)* analisar o agravamento do quadro clínico dos pacientes em duas perspectivas (internação e mortalidade) e desenvolver um modelo de predição para ambas as perspectivas; *ii* avaliar o número de doses de vacinas como um fator de agravamento do quadro clínico; *iii* possibilitar a utilização dos modelos de predição apenas por meio de dados demográficos, histórico de vacinação, sintomas e comorbidades, dispensando o uso dados clínicos; *iv* avaliar um total de 14 algoritmos de aprendizado de máquina durante o desenvolvimento dos modelos; e, *v* disponibilizar os modelos desenvolvidos de modo a possibilitar a sua utilização.

Os modelos foram treinados a partir de dados de pacientes com casos suspeitos de COVID-19 atendidos em hospitais de São Paulo (o estado brasileiro com maior número de instâncias). A partir da etapa de treinamento e teste, foi observado que o algoritmo *Gradient Boosting* obteve os melhores resultados, alcançando uma acurácia de 83% e AUC de 0.89 no modelo de predição da mortalidade, e acurácia de 71% e AUC de 0.75 no modelo de predição da internação. Em seguida, mediante os resultados dos modelos, foi identificado que a idade avançada, falta de ar, ausência de vacinação e número de dias de sintomas foram as principais características atreladas ao agravamento do quadro clínico. Ao final, uma aplicação *Web* foi desenvolvida para possibilitar a utilização dos modelos criados.

No entanto, dentre as limitações apresentadas, é importante destacar que embora esse estudo tenha uma amostra razoável de instâncias, o número de pacientes que foram curados, internados ou vieram a óbito é bastante desbalanceado, o que reduz a quantidade de dados utilizados nas etapas de treinamento, teste e validação. Além disso, ao analisar o modelo de predição do risco de internação, é possível observar que os resultados obtidos na etapa de validação foram baixos, a citar a Acurácia de 71% e as métricas de Precisão (65%), *Recall* (63%) e *F1-Score* (63%) na classificação de pacientes internados, fato esse, que prejudica a possibilidade de aplicação do referido modelo em um contexto real.

Como trabalhos futuros, pretende-se: i) coletar uma versão mais atualizada dos dados na plataforma OpenDataSus; ii) avaliar a utilização da técnica *Transfer Learning*, que permite reutilizar um modelo pré-treinado em um novo problema com um conjunto de dados distinto, e iii) verificar a utilização de outros algoritmos e comitês de aprendizado

de máquina no novo conjunto de dados, tendo em vista que os algoritmos selecionados neste trabalho alcançaram resultado similares.

Referências

ABDULAAL, A. *et al.* Clinical utility and functionality of an artificial intelligence–based app to predict mortality in COVID-19: Mixed methods analysis. *JMIR Formative Research*, JMIR Publications Inc., v. 5, n. 7, p. e27992, jul 2021. Citado na página 90.

ABDULAAL, A. *et al.* Prognostic modeling of COVID-19 using artificial intelligence in the united kingdom: Model development and validation. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 22, n. 8, aug 2020. Citado na página 86.

AKTAR, S. *et al.* Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: Statistical analysis and model development. *JMIR Medical Informatics*, JMIR Publications Inc., v. 9, n. 4, p. e25884, apr 2021. Citado na página 90.

ALI, M. *PyCaret: An open source, low-code machine learning library in Python*. [S.l.], 2020. PyCaret version 1.0. Disponível em: <https://www.pycaret.org>. Citado 2 vezes nas páginas 43 e 55.

ALSMADI, T.; ALQUDAH, N.; NAJADAT, H. Prediction of covid-19 patients states using data mining techniques. In: *2021 International Conference on Information Technology (ICIT)*. [S.l.]: IEEE, 2021. Citado na página 85.

AMARAL, C. *et al.* A molecular test based on RT-LAMP for rapid, sensitive and inexpensive colorimetric detection of SARS-CoV-2 in clinical samples. *Scientific Reports*, Springer Science and Business Media LLC, v. 11, n. 1, ago. 2021. Citado na página 20.

ASSAF, D. *et al.* Utilization of machine-learning models to accurately predict the risk for critical covid-19. *Internal and Emergency Medicine*, 2020. Citado 3 vezes nas páginas 16, 37 e 88.

ATLAM, M. *et al.* Coronavirus disease 2019 (COVID-19): survival analysis using deep learning and cox regression model. *Pattern Analysis and Applications*, Springer Science and Business Media LLC, v. 24, n. 3, p. 993–1005, feb 2021. Citado na página 89.

AZNAR-GIMENO, R. *et al.* A clinical decision web to predict ICU admission or death for patients hospitalised with COVID-19 using machine learning algorithms. *International Journal of Environmental Research and Public Health*, MDPI AG, v. 18, n. 16, p. 8677, aug 2021. Citado 2 vezes nas páginas 37 e 90.

BAJ, J. *et al.* Covid-19: Specific and non-specific clinical manifestations and symptoms: The current state of knowledge. *Journal of Clinical Medicine*, MDPI AG, v. 9, n. 6, p. 1753, jun 2020. Citado na página 72.

BENNETT, M. *et al. Methodology to Create Analysis-Naive Holdout Records as well as Train and Test Records for Machine Learning Analyses in Healthcare*. [S.l.]: arXiv, 2022. Citado na página 26.

BOLOURANI, S. *et al.* A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: Model development and validation. *Journal*

of Medical Internet Research, JMIR Publications Inc., v. 23, n. 2, p. e24246, fev. 2021. Citado na página 88.

BOOTH, A.; ABELS, E.; MCCAFFREY, P. Development of a prognostic model for mortality in covid-19 infection using machine learning. *Modern Pathology*, 2020. Citado 2 vezes nas páginas 72 e 87.

BREIMAN, L. *Machine Learning*, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. Citado na página 29.

BURKOV, A. *The Hundred-Page Machine Learning*. 1th. ed. [S.l.: s.n.], 2019. 160 p. ISBN 978-1999579500. Citado na página 27.

CEN, Y. *et al.* Risk factors for disease progression in patients with mild to moderate coronavirus disease 2019—a multi-centre observational study. *Clinical Microbiology and Infection*, Elsevier BV, v. 26, n. 9, p. 1242–1247, set. 2020. Disponível em: <https://doi.org/10.1016/j.cmi.2020.05.041>. Citado na página 16.

CERRI, R.; CARVALHO, A. C. P. L. F. Aprendizado de máquina: Breve introdução e aplicações. *Cadernos de Ciência & Tecnologia*, v. 34, n. 3, dec 2017. Disponível em: <https://seer.sct.embrapa.br/index.php/cct/article/view/26381>. Citado 2 vezes nas páginas 24 e 25.

CHAWLA, N. *et al.* Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, AI Access Foundation, El Segundo, CA, USA, v. 16, n. 1, p. 321–357, jun 2002. ISSN 1076-9757. Citado na página 42.

CHEN, T.; GUESTRIN, C. XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.]: ACM, 2016. Citado na página 30.

CHEN, Y. *et al.* A multimodality machine learning approach to differentiate severe and nonsevere COVID-19: Model development and validation. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 23, n. 4, p. e23948, abr. 2021. Citado 2 vezes nas páginas 37 e 87.

CHENG, F. *et al.* Using machine learning to predict icu transfer in hospitalized covid-19 patients. *Journal of Clinical Medicine*, 2020. Citado 2 vezes nas páginas 37 e 85.

CHOU, E. *et al.* Clinical features of emergency department patients from early COVID-19 pandemic that predict SARS-CoV-2 infection: Machine-learning approach. *Western Journal of Emergency Medicine*, Western Journal of Emergency Medicine, v. 22, n. 2, mar 2021. Citado na página 90.

CHOWDHURY, M. *et al.* An early warning tool for predicting mortality risk of COVID-19 patients using machine learning. *Cognitive Computation*, Springer Science and Business Media LLC, apr 2021. Citado na página 86.

CHUNG, H. *et al.* Prediction and feature importance analysis for severity of COVID-19 in south korea using artificial intelligence: Model development and validation. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 23, n. 4, p. e27060, abr. 2021. Citado 2 vezes nas páginas 37 e 88.

COBRE, A. *et al.* Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? *Computers in Biology and Medicine*, Elsevier BV, v. 134, p. 104531, jul 2021. Citado na página 90.

COLPO, M. *et al.* Attribute selection based on genetic and classification algorithms in the prediction of hospitalization need of COVID-19 patients. In: *XVII Brazilian Symposium on Information Systems*. [S.l.]: ACM, 2021. Citado na página 85.

DAS, A.; MISHRA, S.; GOPALAN, S. Predicting covid-19 community mortality risk using machine learning and development of an online prognostic tool. *PeerJ*, 2020. Citado 3 vezes nas páginas 37, 72 e 87.

FACELI, K. *et al. Inteligência Artificial: Uma abordagem de Aprendizado de Máquina.* Rio de Janeiro, RJ: Grupo Nacional Editorial, 2011. 378 p. ISBN 978-85-216-1880-5. Citado 3 vezes nas páginas 24, 25 e 28.

FAMIGLINI, L. *et al.* Prediction of ICU admission for COVID-19 patients: a machine learning approach based on complete blood count data. In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. [S.l.]: IEEE, 2021. Citado na página 90.

FERNANDES, F. *et al.* A multipurpose machine learning approach to predict COVID-19 negative prognosis in são paulo, brazil. *Scientific Reports*, Springer Science and Business Media LLC, v. 11, n. 1, feb 2021. Citado na página 86.

FERRARI, D. *et al.* Machine learning in predicting respiratory failure in patients with COVID-19 pneumonia—challenges, strengths, and opportunities in a global health emergency. *PLOS ONE*, Public Library of Science (PLoS), v. 15, n. 11, p. e0239172, nov. 2020. Citado na página 88.

FREUND, Y.; SCHAPIRE, R. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, Elsevier BV, v. 55, n. 1, p. 119–139, aug 1997. Citado na página 29.

GUAN, W. *et al.* Clinical characteristics of coronavirus disease 2019 in china. *New England Journal of Medicine*, Massachusetts Medical Society, v. 382, n. 18, p. 1708–1720, abr. 2020. Disponível em: <https://doi.org/10.1056/nejmoa2002032>. Citado na página 16.

GUAN, X. *et al.* Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study. *Annals of Medicine*, Informa UK Limited, v. 53, n. 1, p. 257–266, jan 2021. Citado na página 91.

GUL, M.; HTUN, Z.; INAYAT, A. Role of fever and ambient temperature in COVID-19. *Expert Review of Respiratory Medicine*, Informa UK Limited, v. 15, n. 2, p. 171–173, set. 2020. Citado na página 72.

GULL, H. *et al.* Severity prediction of COVID-19 patients using machine learning classification algorithms: A case study of small city in pakistan with minimal health facility. In: *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*. [S.l.]: IEEE, 2020. Citado na página 91.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. Estados Unidos, EUA: Elsevier, 2011. 744 p. ISBN 978-0-12-381479-1. Citado 2 vezes nas páginas 19 e 24.

HEALTH, H. *COVID-19 basics: Symptoms, spread and other essential information about the coronavirus and COVID-19*. 2022. Https://www.health.harvard.edu/diseases-and-conditions/covid-19-basics. Citado na página 21.

HOLANDA, W. D. *Covid Mortality Risk Prediction*. [S.l.], 2022. Disponível em: <https://github.com/WallaceHolanda/covid_mortality_risk_prediction>. Citado na página 44.

HOLANDA, W. D.; SILVA, L. C.; SOBRINHO, A. A. C. C. Estratégias preditivas na detecção do agravamento do quadro clínico de pacientes com covid-19: Uma revisão de escopo. *Journal of Health Informatics*, v. 13, n. 4, dez 2021. Citado 2 vezes nas páginas 16 e 17.

HOU, W. *et al.* Machining learning predicts the need for escalated care and mortality in COVID-19 patients from clinical variables. *International Journal of Medical Sciences*, Ivyspring International Publisher, v. 18, n. 8, p. 1739–1745, 2021. Citado 2 vezes nas páginas 37 e 87.

IBGE. *Censo Demográfico - Características da População e dos Domicílios*. 2011. Https://cutt.ly/IXPX6Ca. Citado na página 46.

IKEMURA, K. *et al.* Using automated machine learning to predict the mortality of patients with COVID-19: Prediction model development study. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 23, n. 2, p. e23458, feb 2021. Citado 2 vezes nas páginas 37 e 89.

ISER, B. P. M. *et al.* Definição de caso suspeito da COVID-19: uma revisão narrativa dos sinais e sintomas mais frequentes entre os casos confirmados. *Epidemiologia e Serviços de Saúde*, FapUNIFESP (SciELO), v. 29, n. 3, jun. 2020. Disponível em: <https://doi.org/10.5123/s1679-49742020000300018>. Citado na página 16.

JAKOB, C. *et al.* Prediction of COVID-19 deterioration in high-risk patients at diagnosis: an early warning score for advanced COVID-19 developed by machine learning. *Infection*, Springer Science and Business Media LLC, v. 50, n. 2, p. 359–370, jul 2021. Citado 2 vezes nas páginas 37 e 87.

JOBLIB. *Joblib: running Python functions as pipeline jobs*. [S.l.], 2022. Disponível em: <https://joblib.readthedocs.io/>. Citado na página 71.

KAI, P. *et al.* Systematic mapping studies in software engineering. In: *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*. Swindon, GBR: BCS Learning Development Ltd, 2008. p. 68–77. Citado 2 vezes nas páginas 19 e 31.

KHAN, M. *et al.* Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review. *Expert Systems with Applications*, Elsevier BV, v. 185, p. 115695, dec 2021. Citado na página 16.

KIM, H. *et al.* An easy-to-use machine learning model to predict the prognosis of patients with COVID-19: Retrospective cohort study. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 22, n. 11, p. e24225, nov. 2020. Disponível em: <https://doi.org/10.2196/24225>. Citado 3 vezes nas páginas 37, 73 e 89.

KLUYVER, T. *et al.* Jupyter notebooks – a publishing format for reproducible computational workflows. In: LOIZIDES, F.; SCHMIDT, B. (Ed.). *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. [S.l.], 2016. p. 87 – 90. Citado na página 45.

KO, H. *et al.* An artificial intelligence model to predict the mortality of COVID-19 patients at hospital admission time using routine blood samples: Development and validation of an ensemble model. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 22, n. 12, p. e25442, dez. 2020. Disponível em: <https://doi.org/10.2196/25442>. Citado 2 vezes nas páginas 37 e 87.

KUPELI, I.; YüKSEL, F. Coronavirus (COVID -19), advanced age, and malnutrition: A risky coexistence. *Aging Medicine and Healthcare*, Full Universe Integrated Marketing Ltd, v. 11, n. 4, dez. 2020. Citado na página 72.

LEE, Y. W.; CHOI, J. W.; SHIN, E. Machine learning model for diagnostic method prediction in parasitic disease using clinical information. *Expert Systems with Applications*, Elsevier BV, v. 185, p. 115658, dec 2021. Citado na página 16.

LEMAITRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, v. 18, n. 17, p. 1–5, 2017. Citado na página 52.

LI, S. *et al.* Development and external evaluation of predictions models for mortality of COVID-19 patients using machine learning method. *Neural Computing and Applications*, Springer Science and Business Media LLC, jan. 2021. Citado na página 88.

LUELLEN, E. A machine learning explanation of the pathogen-immune relationship of SARS-CoV-2 (COVID-19), and a model to predict immunity and therapeutic opportunity: A comparative effectiveness research study. *JMIRx Med*, JMIR Publications Inc., v. 1, n. 1, p. e23582, oct 2020. Citado na página 89.

LUNDBERG, S. *SHAP*. [S.l.], 2022. Disponível em: <https://github.com/slundberg/shap>. Citado na página 58.

LUNDBERG, S. M.; LEE, S. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 4768–4777. ISBN 9781510860964. Citado na página 43.

MAHBOUB, B. *et al.* Prediction of COVID-19 hospital length of stay and risk of death using artificial intelligence-based modeling. *Frontiers in Medicine*, Frontiers Media SA, v. 8, maio 2021. Citado 2 vezes nas páginas 37 e 87.

MARCOS, M. *et al.* Development of a severity of disease score and classification model by machine learning for hospitalized COVID-19 patients. *PLOS ONE*, Public Library of Science (PLoS), v. 16, n. 4, p. e0240200, apr 2021. Citado na página 90.

MCKINNEY, W. Data structures for statistical computing in python. In: WALT, S. van der; MILLMAN, J. (Ed.). *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 56–61. Citado na página 45.

MIRANDA, I. *et al.* Machine learning prediction of hospitalization due to COVID-19 based on self-reported symptoms: A study for brazil. In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. [S.l.]: IEEE, 2021. Citado na página 85.

MITCHELL, T. M. *Machine Learning*. New York, NY: McGraw-Hil, 1997. 414 p. ISBN 978-0-07-042807-2. Citado 5 vezes nas páginas 16, 19, 24, 25 e 28.

NORONHA, K. *et al.* Pandemia por covid-19 no brasil: análise da demanda e da oferta de leitos hospitalares e equipamentos de ventilação assistida segundo diferentes cenários. *Cadernos de Saúde Pública*, FapUNIFESP (SciELO), v. 36, n. 6, 2020. Disponível em: <https://doi.org/10.1590/0102-311x00115320>. Citado na página 20.

OLIVEIRA, G. *et al. Estudo sobre o impacto da pandemia da COVID-19 nos custos do setor de saúde*. 2022. 102 p. Https://cutt.ly/mVxQ4WX. Citado na página 16.

OLMEDO, J. *et al.* Machine learning applied to clinical laboratory data in spain for COVID-19 outcome prediction: Model development and validation. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 23, n. 4, p. e26211, apr 2021. Citado na página 89.

PAN, P. *et al.* Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: Model development and validation. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 22, n. 11, p. e23128, nov. 2020. Citado na página 88.

PAPOUTSOGLOU, G. *et al.* Automated machine learning optimizes and accelerates predictive modeling from COVID-19 high throughput datasets. *Scientific Reports*, Springer Science and Business Media LLC, v. 11, n. 1, jul. 2021. Citado 2 vezes nas páginas 37 e 87.

PARBATE, N. *et al.* ICU admission prediction using machine learning for covid-19 patients. In: *2021 International Conference on Communication information and Computing Technology (ICCICT)*. [S.l.]: IEEE, 2021. Citado na página 85.

PARCHURE, P. *et al.* Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with COVID-19. *BMJ Supportive & Palliative Care*, BMJ, p. bmjspcare–2020–002602, set. 2020. Citado na página 89.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in python. *Journal of machine learning research*, v. 12, n. Oct, p. 2825–2830, 2011. Citado na página 71.

PENG, C.; LEE, K.; INGERSOLL, G. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, Informa UK Limited, v. 96, n. 1, p. 3–14, sep 2002. Citado na página 29.

PODDER, P.; MONDAL, M. R. Machine learning to predict COVID-19 and ICU requirement. In: *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*. [S.l.]: IEEE, 2020. Citado na página 85.

POUDEL, A. *et al.* Impact of covid-19 on health-related quality of life of patients: A structured review. *PLOS ONE*, Public Library of Science (PLoS), v. 16, n. 10, p. e0259164, out. 2021. Citado na página 21.

RAMALHO, J. M. *et al.* Nursing diagnosis outcomes and interventions for critically ill patients affected by covid-19 and sepsis. *Texto & Contexto - Enfermagem*, FapUNIFESP (SciELO), v. 29, 2020. Disponível em: <https://doi.org/10.1590/1980-265x-tce-2020-0160>. Citado na página 16.

RAZU, S. *et al.* Challenges faced by healthcare professionals during the COVID-19 pandemic: A qualitative inquiry from bangladesh. *Frontiers in Public Health*, Frontiers Media SA, v. 9, ago. 2021. Citado na página 21.

REZENDE, S. O. *Sistemas Inteligentes: Fundamentos e Aplicações*. Barueri, SP: Editora Manole Ltda, 2003. ISBN 8520416837. Citado na página 26.

ROCHA, F. G.; NASCIMENTO, B. A.; NASCIMENTO, E. F. Um modelo de mapeamento sistemático para a educação. *Cadernos da FUCAMP*, 2018. Disponível em: <https://revistas.fucamp.edu.br/index.php/cadernos/issue/view/86>. Citado na página 31.

SANTANA, Í. V. dos S. *et al.* Classification models for COVID-19 test prioritization in brazil: Machine learning approach. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 23, n. 4, p. e27293, abr. 2021. Disponível em: <https://doi.org/10.2196/27293>. Citado na página 16.

SCHONING, V. *et al.* Development and validation of a prognostic COVID-19 severity assessment (COSA) score and machine learning models for patient triage at a tertiary hospital. *Journal of Translational Medicine*, Springer Science and Business Media LLC, v. 19, n. 1, fev. 2021. Citado na página 88.

SHAO, Y. *et al.* Understanding demographic risk factors for adverse outcomes in COVID-19 patients: Explanation of a deep learning model. *Journal of Healthcare Informatics Research*, Springer Science and Business Media LLC, v. 5, n. 2, p. 181–200, feb 2021. Citado na página 90.

SHEN, J. *et al.* Decision support analysis for risk identification and control of patients affected by COVID-19 based on bayesian networks. *Expert Systems with Applications*, Elsevier BV, v. 196, p. 116547, jun 2022. Disponível em: <https://doi.org/10.1016/j.eswa.2022.116547>. Citado na página 16.

SHI, L. *et al.* Dyspnea rather than fever is a risk factor for predicting mortality in patients with COVID-19. *Journal of Infection*, Elsevier BV, v. 81, n. 4, p. 647–679, out. 2020. Citado na página 72.

SHUKLA, S. K. *et al.* Progress in COVID research and developments during pandemic. *VIEW*, Wiley, p. 20210020, jul. 2022. Citado na página 20.

SIDEY-GIBBONS, J. A. M.; SIDEY-GIBBONS, C. J. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, Springer Science and Business Media LLC, v. 19, n. 1, mar. 2019. Disponível em: <https://doi.org/10.1186/s12874-019-0681-4>. Citado na página 16.

SILVA, L. *et al.* Importance of covid-19 early diagnosis: a literature review. *Brazilian Journal of Health Review*, South Florida Publishing LLC, v. 4, n. 5, p. 23659–23673, nov 2021. Citado na página 16.

SINGHAL, S. *et al.* Clinical features and outcomes of COVID-19 in older adults: a systematic review and meta-analysis. *BMC Geriatrics*, Springer Science and Business Media LLC, v. 21, n. 1, maio 2021. Citado na página 72.

STACHEL, A. *et al.* Development and validation of a machine learning model to predict mortality risk in patients with COVID-19. *BMJ Health &amp Care Informatics*, BMJ, v. 28, n. 1, p. e100235, may 2021. Citado na página 89.

STREAMLIT. *Streamlit: The fastest way to build and share data apps*. [S.l.], 2022. Disponível em: <https://github.com/streamlit/streamlit>. Citado na página 71.

SUBUDHI, S. *et al.* Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *npj Digital Medicine*, Springer Science and Business Media LLC, v. 4, n. 1, may 2021. Citado 2 vezes nas páginas 37 e 86.

SUN, C. *et al.* Predicting COVID-19 disease progression and patient outcomes based on temporal deep learning. *BMC Medical Informatics and Decision Making*, Springer Science and Business Media LLC, v. 21, n. 1, feb 2021. Citado na página 86.

TEYMOURI, M. *et al.* Recent advances and challenges of RT-PCR tests for the diagnosis of COVID-19. *Pathology - Research and Practice*, Elsevier BV, v. 221, p. 153443, maio 2021. Citado na página 20.

TEZZA, F. *et al.* Predicting in-hospital mortality of patients with COVID-19 using machine learning techniques. *Journal of Personalized Medicine*, MDPI AG, v. 11, n. 5, p. 343, apr 2021. Citado na página 86.

VAID, A. *et al.* Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in new york city: Model development and validation. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 22, n. 11, p. e24018, nov. 2020. Disponível em: <https://doi.org/10.2196/24018>. Citado na página 88.

VANDERPLAS, J. *Python Data Science Handbook: Essential Tools for Working with Data*. Estados Unidos, EUA: O'Reilly Media, 2016. 548 p. ISBN 978-1491912058. Citado na página 20.

WANG, R. *et al.* Predictions of COVID-19 infection severity based on co-associations between the SNPs of co-morbid diseases and COVID-19 through machine learning of genetic data. In: *2020 IEEE 8th International Conference on Computer Science and Network Technology (ICCSNT)*. [S.l.]: IEEE, 2020. Citado na página 85.

WANG, W.; LU, Y. Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, v. 324, p. 012049, mar. 2018. Citado na página 26.

WHO. *Coronavirus*. 2020. Https://www.who.int/health-topics/coronavirus#tab=tab_1. Citado na página 16.

WICKHAM, H.; GROLEMUND, G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Canada, CA: O'Reilly Media, 2017. 520 p. ISBN 978-8550803241. Citado na página 23.

WIERINGA, R. J. *Design Science Methodology for Information Systems and Software Engineering*. 1. ed. [S.l.]: Springer-Verlag Berlin Heidelberg, 2014. Citado 2 vezes nas páginas 17 e 19.

YAO, H. *et al.* Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests. *Frontiers in Cell and Developmental Biology*, Frontiers Media SA, v. 8, jul 2020. Citado na página 89.

YE, J.; HUA, M.; ZHU, F. Machine learning algorithms are superior to conventional regression models in predicting risk stratification of COVID-19 patients. *Risk Management and Healthcare Policy*, Informa UK Limited, Volume 14, p. 3159–3166, jul 2021. Citado na página 86.

YU, L. *et al.* Machine learning methods to predict mechanical ventilation and mortality in patients with COVID-19. *PLOS ONE*, Public Library of Science (PLoS), v. 16, n. 4, p. e0249285, apr 2021. Citado na página 86.

ZAYET, S. *et al.* Clinical features of COVID-19 and influenza: a comparative study on nord franche-comte cluster. *Microbes and Infection*, Elsevier BV, v. 22, n. 9, p. 481–488, oct 2020. Citado na página 72.

ZHU, J. *et al.* Deep-learning artificial intelligence analysis of clinical variables predicts mortality in COVID-19 patients. *Journal of the American College of Emergency Physicians Open*, Wiley, v. 1, n. 6, p. 1364–1373, aug 2020. Citado 2 vezes nas páginas 16 e 85.

APÊNDICE A – Trabalhos Selecionados no Mapeamento Sistemático

Autor	Algoritmo	Software	Tipo de Dados	Principais Variáveis	Acesso
Wang *et al.* (2020)	RF	Python	Sinais e Sintomas	Doença Renal Crônica, Doença pulmonar crônica, Doença arterial	Não
Alsmadi, Alqudah e Najadat (2021)	SVM	Não Informado	Demográfico	Não Informado	Não
Miranda *et al.* (2021)	MLP	Não Informado	Demográfico, Sinais e Sintomas	Não Informado	Sim
Parbate *et al.* (2021)	XGBoost	Python	Demográfico, Sinais e Sintomas, Clínico	Não Informado	Não
Colpo *et al.* (2021)	DT	Python	Demográfico, Sinais e Sintomas	Não Informado	Não
Zhu *et al.* (2020)	DNN	R	Demográfico, Sinais e Sintomas	Teste D-Dímero, Índice O2, Proporção de neutrófilos	Não
Podder e Mondal (2020)	Stacking (RF, ET e LR)	Python	Demográfico e Clínico	Proteína C, Idade, Exame SARS-CoV-2	Não
Cheng *et al.* (2020)	RF	R e Apache Spark	Demográfico, Sinais e Sintomas, Clínico	Frequência Respiratória, Contagem de leucócitos, Idade	Não

Subudhi *et al.* (2021)	AdaBoost	R e Python	Demográfico, Sinais e Sintomas, Clínico	Doença Renal Crônica, Porcentagens de neutrófilos, Linfócitos	Não
Tezza *et al.* (2021)	RF	R	Demográfico, Sinais e Sintomas, Clínico	Idade, Saturação, SOFA rápido	Sim
Sun *et al.* (2021)	NN T-LSTM	Python	Demográfico, Clínico	Linfa, LDH, Hs-CRP	Não
Ye, Hua e Zhu (2021)	XGBoost	R	Demográfico, Sinais e Sintomas, Clínico	Proteína C, Procalcitonina; Porcentagem de neutrófilos	Não
Yu *et al.* (2021)	CatBoost	Python	Demográfico, Sinais e Sintomas	Ventilação Mecânica, Respiração, IMC	Não
Abdulaal *et al.* (2020)	NN	Python	Demográfico, Sinais e Sintomas	Mentação alterada, Dispneia, Idade avançada	Não
Fernandes *et al.* (2021)	RF	Não Informado	Demográfico, Clínico	Idade, Proteína C, Creatinina	Não
Chowdhury *et al.* (2021)	XGBoost	R	Demográfico, Clínico	Lactato Desidrogenase, Neutrófilos, Linfócitos	Não

Jakob *et al.* (2021)	XGBoost	R e H2O.ai	Demográfico, Sinais e Sintomas, Clínico	Saturação, Idade, Creatinina	Sim
Das, Mishra e Gopalan (2020)	LR	Python e R	Demográfico	Sexo, Idade, Localidade	Sim
Booth, Abels e McCaffrey (2020)	SVM	Python	Clínico	CRP, Nitrogênio uréico, Cálcio sérico	Não
Papoutsoglou *et al.* (2021)	Não Informado	R e JADBio	Clínico	Não Informado	Não
Hou *et al.* (2021)	RF	R e H2O.ai	Demográfico, Sinais e Sintomas, Clínico	Idade, Procalcitonina, Proteína C	Não
Mahboub *et al.* (2021)	DT	R e H2O.ai	Demográfico, Sinais e Sintomas, Clínico	IMC, Problema renal, Uso de anticoagulante	Não
Chen *et al.* (2021)	RF	Python e R	Demográfico, Sinais e Sintomas, Clínico	Idade, Hipertensão, Doença cardiovascular	Não
Ko *et al.* (2020)	Ensemble (NN e RF)	Python e TensorFlow	Demográfico; Clínico	Linfócitos, Neutrófilos, Monócitos	Sim

Li *et al.* (2021)	XGBoost	Não Informado	Demográfico, Sinais e Sintomas, Clínico	Desidrogenase láctica, nitrogênio Uréico, Linfócitos	Não
Chung *et al.* (2021)	DNN	R e TensorFlow	Demográfico, Sinais e Sintomas, Clínico	Idade, Linfócitos, Contagem de plaquetas	Sim
Ferrari *et al.* (2020)	LightGBM	Não Informado	Demográfico, Sinais e Sintomas, Clínico	Dispnéia, Proteína C, Pressão parcial do sangue	Não
Bolourani *et al.* (2021)	XGBoost	Python	Demográfico, Sinais e Sintomas, Clínico	Fornecimento de oxigênio, Idade, Gravidade de emergência	Não
Schoning *et al.* (2021)	SVM	R	Demográfico, Sinais e Sintomas, Clínico	Sexo, Proteína C, Sódio	Não
Assaf *et al.* (2020)	NN	SPSS 25 e JASP	Demográfico, Sinais e Sintomas, Clínico	Dias de infecção, Linfócitos, Sódio	Não
Vaid *et al.* (2020)	XGBoost	Python	Demográfico, Sinais e Sintomas, Clínico	Idade, Hiato aniônico, Proteína-C	Não
Pan *et al.* (2020)	XGBoost	Python	Sinais e Sintomas, Clínico	Protrombina, Saturação, Linfócitos	Não

Parchure *et al.* (2020)	RF	Apache Spark	Demográfico, Sinais e Sintomas, Clínico	Idade, Estado da Função Renal, Proteína C	Não
Kim *et al.* (2020)	XGBoost	R e H2O.ai	Demográfico, Sinais e Sintomas	Idade, Sexo, Histórico de fumante	Sim
Luellen (2020)	RF	R e Minitab	Clínico	CGF, IL-16, HGF	Não
Yao *et al.* (2020)	SVM	Python	Demográfico, Clínico	Neutrófilos, Cálcio no sangue, Idade	Sim
Atlam *et al.* (2021)	DNN	Não Informado	Demográfico, Sinais e Sintomas, Clínico	Idade, Pneumonia, Dores musculares	Não
Stachel *et al.* (2021)	XGBoost	Python	Demográfico, Sinais e Sintomas, Clínico	Oximetria, Respiração, Nitrogênio ureico	Não
Olmedo *et al.* (2021)	XGBoost	Python	Demográfico, Clínico	Lactato desidrogenase, Proteína C, Neutrófilos	Não
Ikemura *et al.* (2021)	XGBoost	R e H2O.ai	Demográfico, Sinais e Sintomas, Clínico	Pressão arterial, Idade, Oximetria de pulso	Não

Chou *et al.* (2021)	RF	Python	Demográfico, Sinais e Sintomas	Temperatura, Peso, IMC	Não
Cobre *et al.* (2021)	ANN	Não Informado	Clínico	Hiperferritinemia, Hipocalcemia, Hipoxemia	Não
Aktar *et al.* (2021)	LightGBM	Não Informado	Demográfico, Sinais e Sintomas	Lactato, Frequência respiratória, Pressão arterial diastólica	Não
Abdulaal *et al.* (2021)	AAN	Tensorflow.js	Demográfico, Sinais e Sintomas	Não Informado	Sim
Aznar-Gimeno *et al.* (2021)	XGBoost	Python e R	Demográfico, Sinais e Sintomas	Uréia, Idade, Linfócitos	Não
Marcos *et al.* (2021)	LR	Python	Demográfico, Sinais e Sintomas, Clínico	Sturação, Idade, Filtração glomerular	Sim
Shao *et al.* (2021)	MLP	Não Informado	Demográfico, Sinais e Sintomas	Idade, Sexo, Raça	Não
Famiglini *et al.* (2021)	Ensemble (XGBoost, SVM e RF)	Não Informado	Demográfico, Clínico	Neutrófilos, Eosinófilos, Sexo	Não

Guan *et al.* (2021)	XGBoost	Não Informado	Demográfico, Sinais e Sintomas, Clínico	Gravidade da doença, Idade, Hs-CRP	Não
Gull *et al.* (2020)	SVM	Python	Sinais e Sintomas	Febre, Cansaço, Tosse seca	Não

APÊNDICE B – Desempenho dos Algoritmos em relação aos Dados dos Pacientes que vieram a Óbito

Análise das métricas de desempenho dos algoritmos em relação a base de dados 60/40.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ada	Ada Boost Classifier	0.8134	0.8422	0.8931	0.8175	0.8535	0.5978	0.6027	0.268
gbc	Gradient Boosting Classifier	0.8106	0.8423	0.8938	0.8138	0.8518	0.5910	0.5965	0.371
lr	Logistic Regression	0.8098	0.8451	0.8787	0.8216	0.8490	0.5927	0.5958	0.442
lda	Linear Discriminant Analysis	0.8098	0.8452	0.8800	0.8208	0.8492	0.5925	0.5957	0.049
rf	Random Forest Classifier	0.7782	0.8107	0.8603	0.7935	0.8254	0.5229	0.5265	0.794
knn	K Neighbors Classifier	0.7735	0.7937	0.8734	0.7809	0.8244	0.5078	0.5147	0.123
et	Extra Trees Classifier	0.7635	0.7884	0.8334	0.7898	0.8108	0.4957	0.4975	1.030
svm	SVM - Linear Kernel	0.7507	0.0000	0.8375	0.7886	0.7967	0.4621	0.4951	0.030
dt	Decision Tree Classifier	0.7043	0.6990	0.7443	0.7644	0.7540	0.3837	0.3843	0.020
nb	Naive Bayes	0.6688	0.8024	0.9390	0.6606	0.7755	0.2114	0.2683	0.017

Análise das métricas de desempenho dos algoritmos em relação a base de dados 70/30.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.8421	0.8521	0.9346	0.8553	0.8931	0.5931	0.6025	0.167
ada	Ada Boost Classifier	0.8388	0.8499	0.9304	0.8545	0.8907	0.5856	0.5943	0.183
gbc	Gradient Boosting Classifier	0.8367	0.8466	0.9351	0.8493	0.8900	0.5761	0.5875	0.319
rf	Random Forest Classifier	0.8095	0.8171	0.8994	0.8419	0.8695	0.5176	0.5226	0.666
knn	K Neighbors Classifier	0.8080	0.7949	0.9224	0.8265	0.8715	0.4953	0.5090	0.174
et	Extra Trees Classifier	0.8017	0.7906	0.8846	0.8427	0.8630	0.5045	0.5074	0.663
nb	Naive Bayes	0.7729	0.8158	0.7964	0.8725	0.8315	0.4837	0.4915	0.019
dt	Decision Tree Classifier	0.7481	0.7079	0.8142	0.8269	0.8202	0.3988	0.3996	0.028
dummy	Dummy Classifier	0.7061	0.5000	1.0000	0.7061	0.8277	0.0000	0.0000	0.014
qda	Quadratic Discriminant Analysis	0.6156	0.5396	0.7243	0.7329	0.7231	0.0733	0.0748	0.024

Análise das métricas de desempenho dos algoritmos em relação a base de dados SMOTE 60/40.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.7976	0.8565	0.8202	0.7864	0.8024	0.5951	0.5966	0.824
gbc	Gradient Boosting Classifier	0.7943	0.8532	0.8481	0.7673	0.8052	0.5884	0.5926	0.301
ada	Ada Boost Classifier	0.7853	0.8493	0.8150	0.7706	0.7916	0.5704	0.5724	0.267
lr	Logistic Regression	0.7839	0.8501	0.8196	0.7663	0.7916	0.5678	0.5701	0.077
knn	K Neighbors Classifier	0.7713	0.8265	0.7798	0.7695	0.7736	0.5426	0.5440	0.148
svm	SVM - Linear Kernel	0.7660	0.0000	0.8679	0.7283	0.7875	0.5314	0.5515	0.041
dt	Decision Tree Classifier	0.7437	0.7514	0.7135	0.7614	0.7362	0.4874	0.4889	0.028
nb	Naive Bayes	0.6694	0.8072	0.9377	0.6123	0.7405	0.3375	0.3987	0.019
qda	Quadratic Discriminant Analysis	0.5183	0.5185	0.4662	0.5241	0.4650	0.0370	0.0405	0.033
dummy	Dummy Classifier	0.5020	0.5000	1.0000	0.5020	0.6684	0.0000	0.0000	0.016

Análise das métricas de desempenho dos algoritmos em relação a base de dados SMOTE 70/30.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.8258	0.8956	0.8262	0.8286	0.8271	0.6517	0.6521	0.856
et	Extra Trees Classifier	0.8246	0.8798	0.8118	0.8363	0.8235	0.6492	0.6499	0.755
gbc	Gradient Boosting Classifier	0.7937	0.8681	0.8423	0.7710	0.8048	0.5871	0.5902	0.418
dt	Decision Tree Classifier	0.7888	0.8102	0.7380	0.8251	0.7789	0.5781	0.5816	0.054
lr	Logistic Regression	0.7880	0.8617	0.8241	0.7716	0.7966	0.5756	0.5776	0.131
ridge	Ridge Classifier	0.7875	0.0000	0.8317	0.7672	0.7979	0.5747	0.5774	0.035
lda	Linear Discriminant Analysis	0.7873	0.8614	0.8317	0.7669	0.7977	0.5743	0.5770	0.045
ada	Ada Boost Classifier	0.7847	0.8603	0.8143	0.7720	0.7923	0.5692	0.5706	0.219
svm	SVM - Linear Kernel	0.7835	0.0000	0.8457	0.7599	0.7964	0.5664	0.5781	0.112
nb	Naive Bayes	0.7411	0.8109	0.6893	0.7753	0.7272	0.4827	0.4888	0.034

APÊNDICE C – Desempenho dos Algoritmos a partir dos Dados dos Pacientes Internados

Análise das métricas de desempenho dos algoritmos em relação a base de dados 60/40.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.7726	0.8007	0.8843	0.7731	0.8244	0.5055	0.5173	0.159
lda	Linear Discriminant Analysis	0.7530	0.7851	0.8739	0.7567	0.8107	0.4604	0.4718	0.024
lr	Logistic Regression	0.7498	0.7840	0.8713	0.7548	0.8083	0.4534	0.4645	0.322
ridge	Ridge Classifier	0.7498	0.0000	0.8739	0.7532	0.8087	0.4529	0.4648	0.016
ada	Ada Boost Classifier	0.7451	0.7838	0.8648	0.7521	0.8041	0.4442	0.4545	0.127
lightgbm	Light Gradient Boosting Machine	0.7247	0.7549	0.8167	0.7504	0.7813	0.4109	0.4161	0.124
rf	Random Forest Classifier	0.7144	0.7462	0.8167	0.7392	0.7752	0.3858	0.3915	0.529
et	Extra Trees Classifier	0.6987	0.7237	0.7842	0.7354	0.7582	0.3590	0.3622	0.513
nb	Naive Bayes	0.6877	0.7544	0.9506	0.6711	0.7864	0.2645	0.3321	0.016
knn	K Neighbors Classifier	0.6743	0.6917	0.8101	0.6992	0.7502	0.2887	0.2964	0.121

Análise das métricas de desempenho dos algoritmos em relação a base de dados 70/30.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.7795	0.7738	0.9284	0.7928	0.8550	0.4057	0.4306	0.194
lr	Logistic Regression	0.7766	0.7772	0.9175	0.7956	0.8519	0.4055	0.4255	0.337
ada	Ada Boost Classifier	0.7730	0.7736	0.9091	0.7966	0.8488	0.4009	0.4174	0.138
rf	Random Forest Classifier	0.7595	0.7298	0.8931	0.7913	0.8387	0.3715	0.3840	0.564
lightgbm	Light Gradient Boosting Machine	0.7553	0.7342	0.8813	0.7928	0.8345	0.3697	0.3784	0.129
svm	SVM - Linear Kernel	0.7494	0.0000	0.8873	0.7904	0.8312	0.3344	0.3688	0.028
et	Extra Trees Classifier	0.7435	0.7028	0.8603	0.7921	0.8245	0.3507	0.3559	0.558
knn	K Neighbors Classifier	0.7358	0.6816	0.8922	0.7682	0.8253	0.2946	0.3116	0.123
dummy	Dummy Classifier	0.7005	0.5000	1.0000	0.7005	0.8239	0.0000	0.0000	0.012
dt	Decision Tree Classifier	0.6674	0.6098	0.7584	0.7649	0.7610	0.2128	0.2136	0.022

Análise das métricas de desempenho dos algoritmos em relação a base de dados SMOTE 60/40.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.7515	0.8152	0.7771	0.7431	0.7583	0.5028	0.5055	0.178
rf	Random Forest Classifier	0.7397	0.8116	0.7679	0.7309	0.7481	0.4791	0.4812	0.561
lr	Logistic Regression	0.7383	0.8007	0.7693	0.7291	0.7474	0.4764	0.4788	0.341
lda	Linear Discriminant Analysis	0.7383	0.8013	0.7745	0.7266	0.7486	0.4763	0.4792	0.024
ridge	Ridge Classifier	0.7370	0.0000	0.7719	0.7257	0.7470	0.4737	0.4762	0.015
ada	Ada Boost Classifier	0.7351	0.7988	0.7616	0.7282	0.7425	0.4699	0.4731	0.130
knn	K Neighbors Classifier	0.7200	0.7716	0.7303	0.7210	0.7246	0.4399	0.4411	0.125
svm	SVM - Linear Kernel	0.7121	0.0000	0.8381	0.6874	0.7430	0.4226	0.4547	0.025
dt	Decision Tree Classifier	0.6846	0.6863	0.6525	0.7017	0.6753	0.3697	0.3715	0.020
nb	Naive Bayes	0.6032	0.7572	0.9469	0.5647	0.7072	0.2003	0.2756	0.016

Análise das métricas de desempenho dos algoritmos em relação a base de dados SMOTE 70/30.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.7603	0.8279	0.7460	0.7686	0.7564	0.5206	0.5217	0.604
gbc	Gradient Boosting Classifier	0.7316	0.7990	0.7333	0.7314	0.7318	0.4633	0.4640	0.245
lr	Logistic Regression	0.7169	0.7827	0.7097	0.7206	0.7144	0.4339	0.4348	0.074
ada	Ada Boost Classifier	0.7131	0.7815	0.7047	0.7173	0.7100	0.4263	0.4275	0.155
knn	K Neighbors Classifier	0.7085	0.7720	0.6447	0.7384	0.6878	0.4168	0.4207	0.184
dt	Decision Tree Classifier	0.7056	0.7097	0.6532	0.7309	0.6886	0.4111	0.4147	0.025
svm	SVM - Linear Kernel	0.6836	0.0000	0.7450	0.6803	0.6987	0.3670	0.3878	0.037
nb	Naive Bayes	0.5653	0.7313	0.2015	0.7205	0.2723	0.1292	0.1703	0.018
qda	Quadratic Discriminant Analysis	0.5286	0.5287	0.5746	0.5272	0.5449	0.0574	0.0580	0.022
dummy	Dummy Classifier	0.5008	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.013

ANEXO A – Ficha de Registro Pacientes

MINISTÉRIO DA SAÚDE
SECRETARIA DE VIGILÂNCIA EM SAÚDE

Nº

e-SUS Notifica 16/08/2021

FICHA DE NOTIFICAÇÃO DE SG SUSPEITO DE DOENÇA PELO CORONAVÍRUS 2019 – COVID-19 (B34.2)

Definição de caso: Indivíduo com quadro respiratório agudo, caracterizado por pelo menos dois (2) dos seguintes sinais e sintomas: febre (mesmo que referida), calafrios, dor de garganta, dor de cabeça, tosse, coriza, distúrbios olfativos ou distúrbios gustativos.
Em crianças: além dos itens anteriores considera-se também obstrução nasal, na ausência de outro diagnóstico específico.
Em idosos: deve-se considerar também critérios específicos de agravamento como síncope, confusão mental, sonolência excessiva, irritabilidade e inapetência.
Observação: Na suspeita de COVID-19, a febre pode estar ausente e sintomas gastrointestinais (diarreia) podem estar presentes.

IDENTIFICAÇÃO

Município de Notificação: _____		**UF de notificação:** |__|__|	**Data da Notificação:** ___|___|____
Tem CPF? *(Marcar X)* |__|Sim |__|Não	**Estrangeiro:** *(Marcar X)* |__| Sim |__|Não	**Profissional de saúde:** *(Marcar X)* |__| Sim |__|Não	**Profissional de segurança:** *(Marcar X)* |__| Sim |__|Não
CPF: __|__|__|__|__|__|__|__|__|__|__	**CNS:** __|__|__|__|__|__|__|__|__|__|__|__|__|__|__		**Passaporte:** __|__|__|__|__|__|__|__|__
Ocupação (CBO):			
Nome Completo:			
Nome Completo da Mãe:			
Data de nascimento: ____|____|_____	**País de origem:**		
Sexo: *(Marcar X)* |__|Masculino |__|Feminino	**Raça/Cor:** *(Marcar X)* |__| Branca |__| Preta |__|Amarela |__|Parda |__|Ignorado |__|Indígena Se indígena, informar etnia:_____		
É membro de povo ou comunidade tradicional? *(Marcar X)* |__| Sim |__|Não	***Se sim, qual?*** _____		
Estado de residência: |__|__|	**Município de Residência:** _____		**CEP:** __|__|__|__|__ - __|__|__
Logradouro:		**Número:**	**Bairro:**
Complemento:			
Telefone 1:		**Telefone 2:**	
E-mail:			

ESTRATÉGIA E LOCAL DE REALIZAÇÃO DA TESTAGEM

Estratégia: *(Marcar X)*	|__| Diagnóstico assistencial (sintomático)	|__| Busca ativa de assintomático |__| Triagem de população específica
Se busca ativa de assintomático: *(Marcar X)*	|__| Monitoramento de contatos |__| Investigação de surtos |__| Monitoramento de viajantes com risco de VOC (quarentena) |__| Outro: _____	***Se triagem de população específica:*** *(Marcar X)* |__| Trabalhadores de serviços essenciais ou estratégicos |__| Profissionais de saúde |__| Gestantes e puérperas |__| Povos e comunidades tradicionais |__| Outro: _____
Local de realização da testagem: *(Marcar X)*	|__| Serviço de saúde (UBS, hospital, UPA etc.) |__| Farmácia ou drogaria |__| Outro: _____	|__| Local de trabalho |__| Escola
		|__| Aeroporto |__| Domicílio ou comunidade

DADOS CLÍNICOS EPIDEMIOLÓGICOS

Sintomas: *(Marcar X)* \|__\| Assintomático \|__\| Febre \|__\| Dor de Garganta \|__\| Dispneia \|__\| Tosse \|__\| Coriza \|__\| Dor de Cabeça \|__\| Distúrbios gustativos \|__\| Distúrbios olfativos \|__\| Outros_____

Data do início dos sintomas: ___\|___\|____

Condições: *(Marcar X)*
\|__\| Doenças respiratórias crônicas descompensadas \|__\| Doenças renais crônicas em estágio avançado (graus 3, 4 e 5) \|__\| Portador de doenças cromossômicas ou estado de fragilidade imunológica \|__\| Doenças cardíacas crônicas \|__\| Puérpera (até 45 dias do parto) \|__\| Imunossupressão \|__\| Outros_____ \|__\| Diabetes \|__\| Gestante \|__\| Obesidade

Campos preenchidos automaticamente pelo sistema.

Recebeu vacina Covid-19? *(Marcar X)* |__| Sim |__|Não	***Se recebeu vacina Covid-19, informar:***	**Dose**	**Data da vacinação**	**Laboratório produtor da vacina**	**Lote da vacina**
		1ª dose	___|___|____		
		2ª dose	___|___|____		

EXAMES LABORATORIAIS

Tipo de teste	Estado do teste	Data da coleta	Resultado
RT-PCR	|__|Solicitado |__|Coletado |__|Concluído |__|Não Solicitado	___|___|___	|__|Não detectável |__|Detectável |__|Inconclusivo ou Indeterminado
RT-LAMP	|__|Solicitado |__|Coletado |__|Concluído |__|Não Solicitado	___|___|___	|__|Não detectável |__|Detectável |__|Inconclusivo ou Indeterminado
Teste sorológico IgA	|__|Solicitado |__|Coletado |__|Concluído |__|Não Solicitado	___|___|___	|__|Não reagente |__| Reagente |__|Inconclusivo ou Indeterminado
Teste sorológico IgM	|__|Solicitado |__|Coletado |__|Concluído |__|Não Solicitado	___|___|___	|__|Não reagente |__| Reagente |__|Inconclusivo ou Indeterminado
Teste sorológico IgG	|__|Solicitado |__|Coletado |__|Concluído |__|Não Solicitado	___|___|___	|__|Não reagente |__| Reagente |__|Inconclusivo ou Indeterminado
Teste sorológico – anticorpos totais	|__|Solicitado |__|Coletado |__|Concluído |__|Não Solicitado	___|___|___	|__|Não reagente |__| Reagente |__|Inconclusivo ou Indeterminado
Teste rápido de anticorpo IgM	|__|Solicitado |__|Coletado |__|Concluído |__|Não Solicitado	___|___|___	|__|Não reagente |__| Reagente |__|Inconclusivo ou Indeterminado
Teste rápido de anticorpo IgG	|__|Solicitado |__|Coletado |__|Concluído |__|Não Solicitado	___|___|___	|__|Não reagente |__| Reagente |__|Inconclusivo ou Indeterminado

EXAMES LABORATORIAIS					
Tipo de teste	**Estado do teste**	**Data da coleta**	**Fabricante**	**Lote**	**Resultado**
Teste rápido de antígeno	|__| Solicitado |__| Concluído |__| Coletado |__| Não Solicitado	___|___|___			|__|Não reagente |__| Reagente |__|Inconclusivo ou Indeterminado

ENCERRAMENTO		
Evolução do caso: *(Marcar X)*	**Classificação final:** *(Marcar X)*	**Data de encerramento:**
|__|Cancelado |__|Em tratamento domiciliar |__|Cura |__|Internado |__|Internado em UTI |__|Óbito |__|Ignorado	|__| Descartado |__| Confirmado Clínico-Imagem |__| Confirmado Clínico-Epidemiológico |__| Confirmado Por Critério Clínico |__| Confirmado Laboratorial |__| Síndrome Gripal Não Especificada	____|____|_____

INFORMAÇÕES COMPLEMENTARES E OBSERVAÇÕES

RASTREAMENTO DE CONTATOS							
ID do caso Fonte (preenchido automaticamente pelo sistema)	**ID do contato** *(preenchido automaticamente pelo sistema)*	**Nome completo do contato**	**CPF do contato**	**Telefone 1**	**Telefone 2**	**Relação com o caso** *[1] Domiciliar; [2] Familiar (extradomiciliar); [3] Laboral; [4] Escolar; [5] Evento Social; [6] Outros.*	**Data do último contato**
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	
						|__|	