

Vitor Pires Silva e Souza

Previsão de infecção relacionada à assistência à saúde em pacientes adultos de UTI utilizando ferramentas de inteligência artificial

Goiânia

2024



UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Vitor Pires Silva e Souza

Título do trabalho: Previsão de infecção relacionada a assistência à saúde em pacientes adultos de UTI utilizando ferramentas de inteligência artificial

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [x] SIM [] NÃO¹

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.



Documento assinado eletronicamente por **Deborah Silva Alves Fernandes, Professora do Magistério Superior**, em 02/02/2024, às 16:02, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Vitor Pires Silva E Souza, Discente**, em 06/02/2024, às 15:36, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4358966** e o código CRC **FDFFEBF6**.

Referência: Processo nº 23070.058676/2023-69

SEI nº 4358966

Vitor Pires Silva e Souza

Previsão de infecção relacionada à assistência à saúde em pacientes adultos de UTI utilizando ferramentas de inteligência artificial

Trabalho de conclusão de curso apresentado na Escola de Engenharia Elétrica, Mecânica e de Computação como requisito para a conclusão do curso de Engenharia de Computação e obtenção do título de Engenheiro de Computação.

Universidade Federal de Goiás – UFG

Escola de Engenharia Elétrica, Mecânica e de Computação (EMC)

Orientador: Profa. Dra. Deborah Silva Alves Fernandes

Goiânia

2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Souza, Vitor Pires Silva e

Previsão de infecção relacionada à assistência à saúde em pacientes
adultos de UTI utilizando ferramentas de inteligência artificial
[manuscrito] / Vitor Pires Silva e Souza. - 2024.

XIV, 14 f.: il.

Orientador: Profa. Dra. Deborah Silva Alves Fernandes.

Trabalho de Conclusão de Curso (Graduação) - Universidade
Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de
Computação (EMC), Engenharia da Computação, Goiânia, 2024.

Bibliografia.

Inclui siglas, fotografias, abreviaturas, gráfico, tabelas.

1. Previsão de IRAS. 2. Hospital das Clínicas UFG. 3. Aprendizado
de máquina. 4. Aprendizado supervisionado. 5. Previsão hospitalar. I.
Fernandes, Deborah Silva Alves, orient. II. Título.

CDU 519.25



UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Ao(s) 2 dia(s) do mês de fevereiro do ano de 2024 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado “Previsão de infecção relacionada a assistência à saúde em pacientes adultos de UTI utilizando ferramentas de inteligência artificial”, de autoria de Vitor Pires Silva e Souza (matrícula:201802749), do curso de Engenharia de Computação, da Escola de Engenharia Elétrica, Mecânica e de Computação da UFG. Os trabalhos foram instalados pela Profa. Dra. Deborah Silva Alves Fernandes (INF/UFG) com a participação dos demais membros da Banca Examinadora: Prof. Dr. Sandrerley Ramos Pires (EMC/UFG) e Profa. Dra. Silvana de Lima Vieira dos Santos (FEN/UFG). Após a apresentação, a banca examinadora realizou a arguição do estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 10,0, tendo sido o TCC considerado aprovado.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.



Documento assinado eletronicamente por **Deborah Silva Alves Fernandes, Professora do Magistério Superior**, em 07/02/2024, às 16:40, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Silvana De Lima Vieira Dos Santos, Professor do Magistério Superior**, em 07/02/2024, às 16:44, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Sandrerley Ramos Pires, Professor do Magistério Superior**, em 07/02/2024, às 19:51, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4369478** e o código CRC **F29B99FA**.



UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

DECLARAÇÃO

Declaração de Frequência

Eu, profa. Dra. Deborah Silva Alves Fernandes (INF/UFG), como orientadora do Projeto de Final de Curso 2 do discente Vitor Pires Silva e Souza(matrícula 201802749) do curso de Engenharia de Computação da EMC/UFG, declaro que o aluno compareceu em todas as reuniões de acompanhamento realizadas para o projeto (PFC2 de 2023/2), alcançando 100% de frequência. Goiânia, 02 de fevereiro de 2024.



Documento assinado eletronicamente por **Deborah Silva Alves Fernandes, Professora do Magistério Superior**, em 02/02/2024, às 15:44, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4358857** e o código CRC **419DFB1F**.

Referência: Processo nº 23070.058676/2023-69

SEI nº 4358857



FORMULÁRIO DE METADADOS PARA DISPONIBILIZAÇÃO DE TESES EDISSERTAÇÕES NA BDTD/UFG

() Mestrado

() Doutorado

Possui agência de fomento?	() Sim (x) Não	Sigla:	
Título do trabalho:	Previsão de infecção relacionada à assistência à saúde em pacientes adultos de UTI utilizando ferramentas de inteligência artificial		
Título em outro idioma:	Predicting healthcare-associated infection in adult ICU patients using artificial intelligence tools		
Data defesa:	(02/02/2024)		

Nome do autor (a):	Vitor Pires Silva e Souza
Como deseja ser citado:	Vitor Pires Silva e Souza
E-mail:	vitopirs@discente.ufg.br
Link do currículo Lattes:	https://lattes.cnpq.br/1725274303665399

Orientador (a):	Deborah Silva Alves Fernandes
E-mail:	deborah.fernandes@ufg.br
Link do currículo Lattes:	http://lattes.cnpq.br/0380764911708235

Coorientador (a):	
Link do currículo Lattes:	


Nomes dos membros da banca:	
1)	Profa. Dra. Deborah Silva Alves Fernandes (INF/UFG)
2)	Profa. Dra. Silvana De Lima Vieira Dos Santos (FEN/UFG)
3)	Prof. Dr. Sandrerley Ramos Pires (EMC/UFG)
4)	
5)	

Palavras-chave:	Previsão de IRAS, Floresta Aleatória, Aprendizado supervisionado, Previsão hospitalar, Hospital das Clínicas UFG, Aprendizado de máquina
-----------------	--

Palavras-chave (outro idioma):	IRAS Prediction, Random Forest, Supervised Learning, Hospital Forecast, Clinical Hospital UFG, Machine Learning
Nome da unidade acadêmica:	Escola de Engenharia Elétrica, Mecânica e de Computação (EMC)
Programa de Pós-Graduação:	-
Área de conhecimento:	Sistemas Inteligentes

<p>Resumo:</p> <p>Neste estudo os dados de pacientes obtidos no momento de sua admissão no Hospital das Clínicas da UFG, de janeiro à agosto de 2021, foram empregados em um sistema inteligente com o intuito de antecipar a ocorrência de Infecções Relacionadas à Assistência à Saúde (IRAS). Para alcançar esse objetivo, foram empregados diversos algoritmos de classificação, incluindo Random Forest, Decision Tree, Gradient Boosting, Adaboost e XGBoost. A avaliação do desempenho desses modelos foi apresentada por meio da capacidade preditiva com área sob a curva ROC (Receiver Operating Characteristic). Os valores 1,00; 0,96; 0,99; 0,99; 1,00 foram obtidos, respectivamente, para Random Forest, Decision Tree, Adaboost, Gradient Boosting e XGBoost. Tais valores proporcionaram uma visão abrangente sobre a capacidade discriminativa dos modelos. Esses resultados permitiram uma compreensão detalhada da eficácia preditiva dos modelos, fornecendo percepções valiosas para o desenvolvimento de estratégias de prevenção e gestão de IRAS no contexto hospitalar.</p>
<p>Resumo em língua estrangeira:</p> <p>This study proposes to use data from patients from January 2021 to August 2021, obtained at the time of admission to the Hospital das Clínicas UFG, with the aim of anticipating the occurrence of Healthcare-Associated Infections. For this purpose, several classification algorithms were used, including Random Forest, Decision Tree, Gradient Boosting, Adaboost and XGBoost. The performance evaluation of these models was presented through the predictive capacity with an area under the Receiver Operating Characteristic curve of 1.00; 0.97; 0.98; 0.98; 1.00 for Random Forest, Decision Tree, Adaboost, Gradient Boosting and XGBoost, respectively. This process will allow a detailed understanding of the predictive effectiveness of the models, providing valuable insights for the development of Healthcare-Associated Infections prevention and management strategies in the hospital context.</p>


 Profa. Dra. Deborah S. A. Fernandes (INF/UFG)

DocuSigned by:

 2DD218FA686F4A1...

Previsão de infecção relacionada à assistência à saúde em pacientes adultos de UTI utilizando ferramentas de inteligência artificial

Vitor Pires Silva e Souza*, graduando em Engenharia de Computação. Deborah Silva Alves Fernandes†, Professora Associada *EMC/UFG. †INF/UFG. E-mails: vitorpirs@discente.ufg.br*, deborah.fernandes@ufg.br†

Resumo—Neste estudo dados de pacientes obtidos no momento de sua admissão no Hospital das Clínicas da UFG, de janeiro a agosto de 2021, foram empregados em um sistema inteligente com o intuito de antecipar a ocorrência de Infecções Relacionadas à Assistência à Saúde (IRAS). Para alcançar esse objetivo, foram empregados diversos algoritmos de classificação, incluindo *Random Forest*, *Decision Tree*, *Gradient Boosting*, *Adaboost* e *XGBoost*. A avaliação do desempenho desses modelos foi apresentada por meio da capacidade preditiva com área sob a curva ROC (*Receiver Operating Characteristic*). Os valores 1,00; 0,96; 0,99; 0,99; 1,00 foram obtidos, respectivamente, para *Random Forest*, *Decision Tree*, *Adaboost*, *Gradient Boosting* e *XGBoost*. Tais valores proporcionaram uma visão abrangente sobre a capacidade discriminativa dos modelos. Esses resultados permitiram uma compreensão detalhada da eficácia preditiva dos modelos, fornecendo percepções valiosas para o desenvolvimento de estratégias de prevenção e gestão de IRAS no contexto hospitalar.

Palavras-chave—Previsão de IRAS, Floresta Aleatória, Aprendizado supervisionado, Previsão hospitalar, Hospital das Clínicas UFG, Aprendizado de máquina.

Abstract—This study proposes to use data from patients from January 2021 to August 2021, obtained at the time of admission to the Hospital das Clínicas UFG, with the aim of anticipating the occurrence of Healthcare-Associated Infections. For this purpose, several classification algorithms were used, including *Random Forest*, *Decision Tree*, *Gradient Boosting*, *Adaboost* and *XGBoost*. The performance evaluation of these models was presented through the predictive capacity with an area under the *Receiver Operating Characteristic* curve of 1.00; 0.97; 0.98; 0.98; 1.00 for *Random Forest*, *Decision Tree*, *Adaboost*, *Gradient Boosting* and *XGBoost*, respectively. This process will allow a detailed understanding of the predictive effectiveness of the models, providing valuable insights for the development of Healthcare-Associated Infections prevention and management strategies in the hospital context.

Index Terms—IRAS Prediction, Random Forest, Supervised Learning, Hospital Forecast, Clinical Hospital UFG, Machine Learning.

I. INTRODUÇÃO

A importância das ferramentas preditivas no reconhecimento antecipado de Infecções Relacionadas à Assistência à Saúde (IRAS) é destacada pela crescente complexidade do ambiente hospitalar e pelo contínuo esforço em melhorar a qualidade dos cuidados de saúde. A finalidade desta pesquisa é explorar como os algoritmos de aprendizado de máquina podem ser aplicados para prever a incidência do IRAS em pacientes e ajudar os profissionais da saúde na tomada de decisões. Os

dados analisados foram obtidos durante o período entre janeiro e agosto do ano 2021, através das informações fornecidas na admissão dos indivíduos no Hospital das Clínicas da UFG. A atenção será direcionada para implementar e avaliar diferentes algoritmos de classificação como *Random Forest*, *Decision Tree*, *Gradient Boosting* *Adaboost* e *XGBoost* com o intuito de verificar a habilidade discriminativa específica desses modelos. Esses algoritmos foram selecionados por causa da sua habilidade em lidar com conjuntos de dados complexos e não lineares, tendo a possibilidade de identificar padrões sutis que podem ser indicadores das próximas IRAS. A métrica utilizada para avaliar o desempenho dos modelos de classificação é a área AUC (*Area Under the Curve*) sob a curva ROC (*Receiver Operating Characteristic*), que permite mensurar sua capacidade preditiva com precisão. Não apenas proporcionando uma visão abrangente da capacidade preditiva de cada algoritmo, os resultados obtidos também serão fonte valiosa para o desenvolvimento de estratégias eficientes na prevenção e gestão das IRAS em um contexto hospitalar.

Dessa maneira, foi feita uma análise utilizando um conjunto de dados que compreenderam 792 entradas representando os pacientes. Entre esses pacientes internados, 57 deles foram afetados por Infecções Relacionadas à Assistência à Saúde (IRAS). Assim, o estudo foi feito em colaboração com o Núcleo de Estudos e Pesquisas em Enfermagem para Prevenção e Controle de Infecções Relacionadas à Assistência à Saúde (NEPIH) através do projeto de pesquisa intitulado "Indicadores Epidemiológicos de Infecções Relacionadas à Assistência à Saúde (IRAS) no município de Goiânia, Goiás". O projeto foi aprovado pelo Comitê de Ética em Pesquisa (protocolo de aprovação número 1269485/2015). Adicionalmente, registraram-se 161 óbitos entre os pacientes, contribuindo para um cenário de grande desafio. Antes de explorar detalhes mais específicos dos dados, destaca-se a sensibilidade dessas informações, essa métrica avalia a capacidade do método de detectar com sucesso resultados classificados como positivos. É importante ressaltar que a entrega desses dados foi realizada gentilmente pela Dra. Silvana de Lima Vieira, e medidas rigorosas de confidencialidade estão sendo aplicadas. Essa colaboração permitiu uma análise mais aprofundada da diversidade de condições de saúde presentes no conjunto de dados, evidenciando a presença de pacientes diagnosticados com a Covid-19.

A complexidade das complicações associadas à assistência à saúde é refletida pela variedade de tipos de IRAS identificadas nos registros. Como discutido no artigo Teixeira et al. [1],

dentre os tipos de doenças infecciosas mais documentados nos hospitais estão os casos de pneumonia decorrentes da utilização de ventilação mecânica, bem como as infecções primárias no sangue detectáveis em laboratório ou não, além das inflamações nas vias respiratórias e nos órgãos que compõem o sistema urinário.

Com a implementação de técnicas de análise de dados, este estudo pretende colaborar com a melhoria contínua da segurança e qualidade no atendimento ao paciente, fornecendo uma abordagem inovadora na prevenção e antecipação das complicações relacionadas às IRAS.

Este artigo apresentará as técnicas empregadas na Seção 2, um levantamento de trabalhos realizados na área será descrito na Seção 3. Na Seção 4, exibirá o experimento, descrevendo os métodos utilizados, a preparação dos dados, a seleção de características, assim como o treinamento e validação do modelo. Na Seção 5, mostrará os resultados obtidos e uma discussão sobre eles. Por fim, na Seção 6, concluirá o artigo com uma análise geral do projeto realizado.

II. REFERENCIAL TEÓRICO

Este estudo envolve a área de aprendizado de máquina que trata de técnicas computacionais capazes de se adaptar com base em experiências. Essa forma de aprendizado é aplicável a diversas tarefas, incluindo classificação e regressão. Entre os algoritmos de aprendizado de máquina mais empregados, destacam-se o *Random Forest* (RF), *Decision Tree* (DT), *Adaboost* (AB), *Gradient Boosting* (GB) e *XGBoost* (XB).

RF é um algoritmo de aprendizado de máquina que se destaca na classificação e regressão. Ele é construído com base no conceito de aprendizado em conjunto, onde várias árvores de decisão são combinadas para formar um modelo preciso. O processo de construção de um RF pode ser resumido em alguns passos importantes. Primeiro, é preciso coletar um conjunto representativo de dados do problema em questão. Em seguida, o algoritmo cria várias árvores de decisão, cada uma treinada em subconjuntos aleatórios diferentes do conjunto original de dados. Durante o treinamento de cada árvore, uma amostra aleatória das características é considerada a cada divisão, introduzindo um nível extra de aleatoriedade. Depois disso, as previsões feitas por cada árvore são combinadas através do voto (no caso da classificação) ou pela média (no caso da regressão) para gerar a saída final do modelo. Essa abordagem conjunta confere ao RF uma maior robustez, reduzindo o superajuste e aumentando a precisão preditiva, o que o torna uma escolha popular em diversas aplicações de aprendizado de máquina, como visto no site Und [2].

DT é uma estrutura gráfica usada para representar escolhas tomadas após avaliação de diferentes características. Essa estrutura é formada através da divisão recursiva do conjunto de dados em subconjuntos cada vez mais homogêneos. Uma condição de teste é representada por um nó interno, e uma decisão ou previsão são representadas por folhas. Em contraste com o RF, uma técnica de aprendizagem automática que usa várias árvores de decisão em sua fundação, cada uma delas mesmas também alimenta diretamente o método. Dessa maneira, o *overfitting* é um problema comum em modelos de

aprendizado de máquina e o RF foi desenvolvido, em parte, para lidar com essa questão. O *overfitting* ocorre quando um modelo se ajusta muito bem aos dados de treinamento, capturando padrões específicos desse conjunto, mas não conseguindo generalizar adequadamente para novos dados. Enquanto uma única árvore de decisão pode ser propensa a *overfitting*, o RF supera essa questão ao utilizar múltiplas estimativas, como falado no artigo [3].

XB, AB e GB pertencem à família de métodos *ensemble*, mas são diferentes em suas abordagens e características específicas. O XB (*eXtreme Gradient Boosting*) é um modelo aprimorado de gradiente impulsionado, com melhorias em termos de velocidade computacional. Isso ocorre porque o algoritmo usa técnicas como regularização e paralelismo. Já o AB, concentra-se no ajuste dos pesos associados às instâncias do conjunto de dados para dar mais importância àquelas classificadas de maneira errada pelo modelo fraco em etapas passadas. Treinando os modelos fracos, ele consegue juntar suas previsões e criar um modelo robusto. E por fim, o GB é uma abordagem comum que desenvolve sucessivas iterações de modelos para mitigar as imprecisões do modelo prévio. Uma abordagem mais ampla é representada pelas implementações do XB e do AB. Enquanto o XB e o AB têm mecanismos adicionais para combater o *overfitting*, é possível que seja um desafio maior no caso do GB.

Para o experimento a ser relatado na Seção IV, na fase de treinamento, foi utilizado o *K-Fold Cross Validation* que consiste em uma técnica para a avaliação de modelos de aprendizagem de máquina. Durante o treinamento e validação, o conjunto de dados é dividido em k subconjuntos (partes). O modelo é treinado k vezes, cada vez usando k-1 partes como conjunto de treinamento e a parte restante como conjunto de validação. Isso é realizado de forma iterativa, garantindo que cada parte seja utilizada como conjunto de validação exatamente uma vez. Ao final do processo, as métricas de desempenho são médias, fornecendo uma avaliação mais robusta e representativa do modelo em diferentes partes dos dados.

Para visualizar os resultados dos modelos foi adotada a matriz de confusão. Esta desempenha um papel na avaliação dos modelos de classificação em aprendizado de máquina. Comparando as previsões do modelo com as classes reais dos dados, ela oferece uma visão detalhada do desempenho.

Fig. 1. Performance da matriz de confusão.

		Valor Predito	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Na figura 1, verdadeiro positivo (VP) representa os casos em que o modelo acertou ao identificar instâncias pertencentes

à classe positiva. verdadeiro negativo (VN) indica os casos em que o modelo acertou ao identificar instâncias pertencentes à classe negativa. falso positivo (FP) refere-se aos casos em que o modelo errou ao indicar que uma instância pertencia à classe positiva quando, na verdade, pertencia à classe negativa. falso negativo (FN) indica os casos em que o modelo errou ao indicar que uma instância pertencia à classe negativa quando, na verdade, pertencia à classe positiva.

Também como validação, o uso da Curva ROC auxilia na análise do desempenho dos modelos de classificação binária em diversos pontos. A representação da curva representa a taxa de verdadeiros positivos (sensibilidade) em função da taxa de falsos positivos (1 - especificidade). Para quantificar o desempenho discriminatório do modelo, muitas vezes se calcula a área sob a curva ROC (AUC-ROC). Quando o valor da AUC-ROC se aproxima de 1 é possível inferir que o modelo apresenta alto nível e eficiência no seu funcionamento. Por outro lado, quando esse valor está próximo do ponto médio (0.5) significa que sua performance é parecida com aquela alcançada pelo acaso ou pela aleatoriedade em geral. Em situações de desequilíbrio nas classes, ter acesso à curva ROC se torna especialmente relevante pois ela fornece uma visão detalhada da capacidade do modelo para distinguir entre os casos positivos e negativos. As Equações 1 e 2 representam, respectivamente, as fórmulas para sensibilidade (S) e especificidade (E).

$$S = \frac{\text{Verdadeiro.positivos}}{\text{Verdadeiro.positivos} + \text{Falso.negativos}} \quad (1)$$

$$E = \frac{\text{Verdadeiro.negativos}}{\text{Verdadeiro.negativos} + \text{Falso.positivos}} \quad (2)$$

III. REVISÃO BIBLIOGRÁFICA

A. Trabalhos relacionados com Covid-19

O artigo Silva [7] aborda o estudo da Síndrome Respiratória Aguda Grave (SRAG), incluindo o Covid-19, com foco em variáveis clínicas e demográficas que influenciam sua evolução, muitas vezes levando ao óbito. O objetivo foi prever a evolução dos casos de SRAG usando dois conjuntos de variáveis preditoras: clínicas e clínicas-demográficas. O pré-processamento dos dados incluiu a variável "evolucao", onde "1" indicou a recuperação e "2" o óbito. Dados ineficientes para classificação foram removidos e em seguida todas as informações foram categorizadas e dicotomizadas em variáveis binárias, separando as clínicas das demográficas. A divisão em treino (70%) e teste (30%) foi realizada para avaliar o desempenho dos classificadores. Os classificadores utilizados foram o KNN(K-Nearest Neighbors), *Decision Tree*, *Logistic Regression*, *Random Forest*, *Neural Network* e o *XGBoost*. As métricas de precisão, acurácia, *F1-score* e AUC-ROC foram utilizadas para avaliar os resultados dos modelos. O *XGBoost* se destacou ao apresentar melhores resultados em todas as métricas. Além disso, a análise da *Decision Tree* permitiu identificar as 20 variáveis mais importantes para classificação dos dados, revelando fatores cruciais que influenciam a evolução dos casos de SRAG, principalmente os de Covid-19. Dentre esses fatores estão o uso de suporte ventilatório, que mostrou

uma alta taxa de óbito (74% dos pacientes que receberam esse tratamento foram a óbito), internação, idade e região de habitação. Adicionalmente, as variáveis demográficas indicaram que a região Nordeste e, em seguida, a região Norte, foram as mais relevantes.

Ainda sobre o vírus da pandemia, o artigo Freire et al. [5] aborda a demora na detecção do SARS-CoV-2, devido ao alto custo e tempo de realização da reação em cadeia da polimerase (PCR), bem como à complexidade na interpretação dos resultados. O estudo utilizou dados de pacientes reais que forneceram amostras para o teste de SARS-CoV-2 e foram submetidos a testes laboratoriais adicionais durante uma visita ao hospital. Após a limpeza e melhoria das informações, o conjunto de dados foi dividido aleatoriamente em 80% para treinamento e 20% para testes. A *Random Forest* foi utilizada para selecionar as 15 características mais importantes, e os modelos SVM, MLP e RF foram treinados com essas características para classificar novas amostras em positivas (Classe 1) ou negativas (Classe 0) para Covid-19. As métricas de precisão, *Recall* e *F1-score* foram usadas para avaliar o desempenho dos modelos. O modelo RF obteve resultados assertivos para todos os casos positivos e negativos de Covid-19. O modelo foi validado com *k-Fold Cross-Validation*, e a classificação de novas amostras foi realizada em três classes de unidades de internação (enfermaria, unidade semi-intensiva e unidade de terapia intensiva) usando as 15 características mais importantes. Os algoritmos foram executados 30 vezes, resultando em 90 cenários diferentes, onde as previsões mostraram-se assertivas para a maioria dos casos positivos e negativos reais para admissão nas diferentes unidades de internação. Em alguns casos, o modelo MLP previu corretamente todos os casos negativos para admissão na enfermaria.

Em seu estudo, Vianna and de Amorim Busana [9] analisa casos incidência diária de Covid-19 nos municípios de Santa Catarina. No processo de pré-processamento, os dados foram tabulados para representar a incidência dos casos de Covid-19 em cada município, correlacionando-os com a data de resultado do exame de confirmação da doença. A autocorrelação dos dados da série temporal foi realizada por meio de funções de autocorrelação total e parcial. Posteriormente, prosseguiu-se com a modelagem dos dados e a avaliação dos modelos, empregando o método de validação *Walk-Forward* e utilizando uma janela de entrada de dados definida em 7 dias através do processo de *Sliding Window*. Com o intuito de otimizar o desempenho dos modelos, foi aplicado o método *Gridsearching* para ajustar os hiperparâmetros do algoritmo de aprendizagem. A modelagem dos dados foi conduzida utilizando-se uma *Neural Network* recorrente, empregando camadas *Long Short-Term Memory* (LSTM). O desempenho do modelo foi avaliado pela métrica de avaliação RMSE (*Root Mean Squared Error*), que representa a raiz quadrada do erro médio quadrático. Mediante essa abordagem, foi possível realizar a predição da incidência diária de Covid-19 nos municípios de Santa Catarina, com um horizonte de tempo de predição de 14 dias.

No artigo de Pantaleão et al. [11] é avaliada a eficácia do uso de *aprendizado de máquina* (ML) em dados sobre a saúde, previsão de óbito, internação em UTI e necessidade de ventilação mecânica (VM) em pacientes hospitalizados com

Tabela I. Resumo das referências bibliográficas.

Referência	Tipos de Dados	Técnica adotada	Resultados	Problema resolvido
Pessoa et al. [4]	Pacientes admitidos na Unidade de Terapia Intensiva (UTI).	<i>Decision Tree, Random Forest, Adaboost, Gradient Boosting e XGBoost</i> , para a construção de um modelo. O método de validação utilizado foi o <i>K-Fold Cross Validation</i> . As métricas de avaliação para analisar o desempenho dos modelos foram <i>Recall</i> , <i>precisão</i> e <i>AUC-ROC</i> .	AUC-ROC de 0,979; 0,999; 0,980; 0,998 e 1,00 para os algoritmos <i>Decision Tree, Random Forest, Adaboost, Gradient Boosting e XGBoost</i> , respectivamente.	Demonstrou uma capacidade de previsão elevada para choque séptico e hipovolêmico a partir do momento da admissão de pacientes na unidade de terapia intensiva.
Freire et al. [5]	Pacientes atendidos no Hospital Israelita Albert Einstein.	<i>Random Forest (RF)</i> , <i>MLP(Multi Layer Perceptron)</i> e <i>SVM(Support Vector Machine)</i> . Utilizou-se o método de busca em grade (<i>Gridsearching</i>) combinado com <i>K-Fold Cross Validation</i> .	Os três algoritmos utilizados alcançaram uma acurácia superior a 99% na primeira tarefa, que consiste em classificar os resultados do teste para pacientes com suspeita de Covid-19. Além disso, na segunda tarefa, que classifica as unidades de internação de pacientes com Covid-19, a acurácia foi superior a 92%.	Desempenho excepcionalmente alto na identificação e classificação de pacientes com suspeita ou diagnóstico confirmado de Covid-19 e triagem de pacientes por unidade de internação.
Pinheiro et al. [6]	Pacientes do sexo feminino que foram diagnosticadas com câncer de mama.	<i>RSF(Random Survival Forest)</i> . Como também, o índice de concordância <i>C-index</i> e o <i>Integrated Brier Score (IBS)</i> .	Idade, tamanho do tumor e estadiamento como fatores de risco relevantes para análise de sobrevivência. Esses atributos foram identificados como tendo maior ganho de informação pelo algoritmo <i>Random Survival Forest</i> e foram levados em conta na análise de dados.	As informações obtidas pelo algoritmo estão alinhadas com as análises tradicionais e podem ser úteis na identificação e compreensão dos fatores de risco associados a essa doença.
Silva [7]	Dados clínicos de pacientes e dados demográficos.	<i>KNN(K-Nearest Neighbors)</i> , <i>Decision Tree</i> , <i>Logistic Regression</i> , <i>Random Forest</i> , <i>Neural Network</i> e o <i>XGBoost</i> . Para validação é utilizado a acurácia, <i>precisão</i> , <i>sensibilidade</i> , <i>F1-score</i> e a curva <i>AUC-ROC</i> .	<i>XGBoost</i> obteve o melhor resultado na métrica de AUC com o valor de 0,86. Os principais fatores identificados para a classificação foram o uso de suporte ventilatório invasivo, idade acima de 70 anos e internação na Unidade de Terapia Intensiva (UTI). Além disso, os principais atributos demográficos foram morar nas regiões nordeste e norte.	Identificado fatores que contribuem para a evolução dos casos de Síndrome Respiratória Aguda Grave (SRAG), especialmente casos de Covid-19.
Oliveira et al. [8]	Pacientes que diagnosticados com doenças como a tuberculose.	<i>DP(Deep Learning)</i> : uma DL totalmente conectada e uma <i>CNN(Convolutional Neural Network)</i> . Aplicado a técnica de <i>Gridsearch</i> e para validação foi utilizado <i>acurácia</i> , <i>precisão</i> , <i>sensibilidade</i> , <i>especificidade</i> e <i>F1-score</i> .	Foi possível observar que uma das configurações alcançou o maior número de acertos para a classe de óbito, com um total de 238 acertos e com 76,6 % de sensibilidade.	Bom desempenho na previsão dos casos de óbito em relação aos demais testados para todas configurações do modelo.
Vianna and de Amorim Busana [9]	Dados clínicos de pacientes e seus sintomas.	<i>Walk-forward</i> com o processo de <i>sliding window</i> , utilizando uma janela de 7 dias para a entrada dos dados. Utilizado uma <i>Neural Network</i> recorrente com múltiplas camadas <i>Long Short-Term Memory (LSTM)</i> . A métrica de avaliação usada para mensurar a qualidade dos modelos foi o <i>Root Mean Squared Error (RMSE)</i> .	A análise das funções de autocorrelação total e parcial confirmou a existência de uma relação na sequência temporal dos dados, indicando a presença de padrões e dependências na evolução da doença ao longo do tempo.	Possibilitou a previsão da incidência diária de Covid-19 em municípios de Santa Catarina, com um horizonte de tempo de predição de 14 dias.
Coelho et al. [10]	Pacientes internados em três Unidades de Terapia Intensiva (UTIs) de um hospital de Belo Horizonte.	<i>LightGBM</i> . O modelo foi avaliado utilizando a métrica <i>AUC-ROC</i> e o processo de validação cruzada de <i>5-Fold Cross-Validation</i> .	O desempenho da predição alcança valores mais altos quando a robustez e a estabilidade da explicação são maiores. Os modelos de previsão de mortalidade alcançaram valores <i>AUC-ROC</i> de até 0,84 no processo de <i>cross-validation</i> .	Capacidade de prever precocemente se complicações ocorrerão durante o período de internação do paciente na UTI.
Pantaleão et al. [11]	Dados de artigos sobre <i>aprendizado de máquina</i> em sintomas de Covid-19.	<i>Checklist de Downs and Black</i> .	<i>Deep Neural Network</i> obteve a melhor especificidade, com um valor de 0,9563. Por outro lado, a tecnologia <i>XGBoost</i> alcançou a melhor acurácia e AUC, com um valor de 1,0.	Apresentaram bons resultados ao prever prognósticos de morte, necessidade de ventilação mecânica e internação em UTI em pacientes hospitalizados com Covid-19.
Eisenlohr [12]	Pacientes e seus sintomas, incluindo o SARS-CoV-2.	<i>Decision Tree</i> , <i>Random Forest</i> , <i>XGBoost</i> e <i>Adaboost</i> . Utilizou-se os valores das métricas de <i>acurácia</i> , <i>AUC-ROC</i> , <i>Recall</i> , <i>precisão</i> e <i>F1-score</i> de cada modelo para validar o desempenho de cada algoritmo.	<i>Decision Tree</i> SMOTE obteve a melhor métrica <i>AUC-ROC</i> com 97,44%. Entretanto, o modelo <i>Adaboost undersampling</i> , apesar de ter uma pontuação <i>AUC-ROC</i> ligeiramente inferior (96,97%), apresentou um desempenho significativamente melhor em relação às métricas de <i>precisão</i> e <i>F1-score</i> .	Classificação da maioria dos casos em que os pacientes faleceram, mas os modelos fizeram várias previsões incorretas de óbito para pacientes que, na realidade, sobreviveram à infecção.
Assaf et al. [13]	Pacientes com Covid-19.	<i>Neural Network</i> , <i>Random Forest</i> e <i>CRT(Classification And Regression Trees)</i> . Modelos avaliados usando métricas como <i>média</i> , <i>sensibilidade</i> , <i>especificidade</i> , <i>valor preditivo positivo</i> , <i>precisão</i> e <i>traçadas curvas ROC</i> e calculada a área média sob a curva <i>AUC</i> para cada gráfico.	<i>Random Forest (RF)</i> apresentou uma melhora de 12,0% em relação ao <i>APACHE II</i> score na predição de resultados clínicos. O modelo <i>CRT</i> alcançou <i>sensibilidade</i> , <i>especificidade</i> , <i>valor preditivo positivo (PPV)</i> , <i>valor preditivo negativo (NPV)</i> e <i>precisão</i> de 88,0%, 92,7%, 68,8%, 97,7% e 92,0%, respectivamente, com uma <i>AUC-ROC</i> de 0,90.	A precisão diagnóstica e a eficácia são aumentadas ao prever o risco de pacientes com Covid-19 desenvolverem doença grave durante o curso da infecção.
Prakash et al. [14]	Pessoas com Covid-19 da Índia e outro contendo dados mundiais de Covid-19.	<i>Random Forest Regressor</i> , <i>Random Forest Classifier</i> , <i>SVM (Support Vector Machine)</i> , <i>Decision Tree</i> , <i>Logistic Regression</i> , <i>XGBoost</i> , <i>Multilinear Regression</i> , <i>GNB Classifier</i> , <i>KNN</i> , <i>RSquared</i> (coeficiente de determinação) e a <i>precisão</i> são calculadas para validar o desempenho de cada modelo.	<i>Random Forest Regressor</i> e o <i>Random Forest Classifier</i> superaram outros modelos em termos de Coeficiente de Determinação (CoD) e <i>precisão</i> . Dessa maneira, <i>RF Regressor</i> e <i>RF Classifier</i> chegaram a atingir 0,97 e 0,92 de CoD, respectivamente.	Foi verificado que pessoas nas faixas etárias de 20 a 29, 30 a 39 e 40 a 50 anos são mais afetadas pelo Covid-19, conforme identificado pelos modelos de previsão.
Schmidt et al. [15]	Sintomas de pacientes do <i>Multiparameter Intelligent Monitoring in Intensive Care III (MIMIC III)</i> .	<i>Neural Network Convolutional (RNC)</i> com cinco estágios e nove camadas ocultas. Para avaliar o desempenho dos modelos, foram utilizadas medidas de <i>acurácia</i> , <i>sensibilidade</i> , <i>especificidade</i> e <i>AUC-ROC</i> .	O modelo proposto com o <i>APACHE II</i> , observou-se que o modelo proposto apresentou bons resultados, atingindo uma área sob a curva <i>ROC (AUC)</i> acima de 0,80. Em contraste, o <i>APACHE II</i> alcançou uma <i>AUC</i> de 0,71.	Modelos propostos possuem uma capacidade semelhante ao <i>APACHE II</i> score na identificação de casos positivos (<i>sensibilidade</i>).
Krishnamoorthi et al. [16]	Pacientes diagnosticados com diabetes.	Foi feito ajuste de hiperparâmetros e a validação cruzada e aplicados nos algoritmos <i>KNN</i> , <i>LR (Logistic Regression)</i> , <i>SVM (Support Vector Machine)</i> e <i>RF (Random Forest)</i> . Assim, é utilizado métricas de <i>sensibilidade</i> , <i>Recall</i> , <i>precisão</i> e <i>acurácia</i> .	<i>Logistic Regression (LR)</i> teve um melhor desempenho em comparação com outros algoritmos de <i>aprendizado de máquina</i> . Os resultados também revelam uma forte correlação entre os níveis de glicose e o índice de massa corporal (IMC) com o diabetes. Além disso, o valor da área sob a curva <i>ROC (AUC)</i> para o algoritmo de <i>Logistic Regression</i> foi de 86%.	Potencial do modelo para auxiliar no diagnóstico precoce e previsão de diabetes, proporcionando uma abordagem promissora para o tratamento dos pacientes.
Shravya et al. [17]	Conjunto de dados retirado do repositório do UCI (<i>University of California Irvine</i>) de pacientes com câncer de mama e seus sintomas.	<i>Logistic Regression</i> , <i>Support Vector Machine (SVM)</i> e <i>K-Nearest Neighbors (KNN)</i> . Foram avaliados resultados de <i>acurácia</i> , <i>precisão</i> e <i>sensibilidade</i> , <i>especificidade</i> .	Dos três algoritmos aplicados, o <i>SVM</i> obteve a maior <i>precisão</i> , alcançando 92,7%.	Prever a ocorrência de câncer de mama com esses conjuntos de dados complexos, fornecendo ferramentas promissoras para tratamento médico.
Santos et al. [18]	Pacientes idosos residentes no Município de São Paulo, Brasil, participantes do estudo Saúde Bem-estar e Envelhecimento (SABE).	<i>Logistic Regression</i> com e sem penalização, <i>Neural Network</i> , <i>Gradient Boosted Trees</i> e <i>Random Forest</i> . Utilizado validação cruzada de 10 campos e a métrica utilizada para seleção dos modelos foi a <i>AUC-ROC</i> .	Apresentaram um desempenho satisfatório, com <i>AUC-ROC</i> superior a 0,70 para todos os modelos propostos.	Previsão de casos futuros de óbito em idosos com base em seus sintomas: um modelo para identificação de riscos.

Covid-19. Para isso, foram utilizados diversos trabalhos com as técnicas utilizadas, métodos de avaliação e resultados obtidos utilizando inteligência artificial. Assim, os estudos foram agrupados com base no prognóstico avaliado, permitindo uma análise detalhada. No que diz respeito ao desfecho de internação em UTI, todos os métodos demonstraram sensibilidade igual ou superior a 0,72, especificidade igual ou superior a 0,709 e

acurácia igual ou superior a 0,728. Esses resultados indicam que os estudos analisados alcançaram uma alta qualidade em suas predições, com pelo menos 60% da pontuação total sendo atingida.

Como em Vianna and de Amorim Busana [9], o estudo realizado por Eisenlohr [12] também realizou a classificação de novos pacientes com base na probabilidade de sobrevivência

à infecção pelo novo coronavírus. Para lidar com o desbalanceamento dos dados, empregaram técnicas de *undersampling* e *oversampling* no pré-processamento. Em seguida, os modelos foram aplicados, incluindo *Random Forest*, *Decision Tree*, *Adaboost*, entre outros. A avaliação do desempenho desses modelos utilizou métricas como acurácia, precisão, *F1-score*, *Recall* e AUC-ROC. Focando especificamente na análise de sobrevivência ou óbito pela Covid-19, os pesquisadores criaram uma nova variável denominada "SobreviveuCovid", que considerava apenas os valores relacionados à "cura" ou "óbito pelo Covid-19". Os autores identificaram os melhores hiperparâmetros e técnicas de balanceamento de dados para cada algoritmo, e os modelos foram novamente treinados com a inclusão dos dados de validação usados anteriormente para testes. Posteriormente, buscaram compreender quais características dos pacientes mais influenciaram suas chances de óbito, sendo a faixa etária elevada, com idade superior a 60 anos.

Na pesquisa realizada por Assaf et al. [13] foram utilizados como critérios a baixa saturação de oxigênio e baixa pressão arterial parcial de oxigênio para prever infecções graves por Covid-19 em pacientes. Para alcançar esse objetivo, foram utilizados três algoritmos de aprendizado de máquina distintos: *Neural Network*, *Random Forest* e *Classification and Regression Trees* (CRT). O desempenho dos modelos foi avaliado por meio de métricas como média, sensibilidade, especificidade, valor preditivo positivo e precisão. Além disso, foram plotadas as curvas características do operador receptor AUC-ROC para cada gráfico. Ao analisar a capacidade de diagnóstico de cada parâmetro isolado utilizando a análise ROC, constataram que a saturação de oxigênio no ar ambiente foi o melhor preditor isolado, com uma AUC de 0,787 ($p < 0,001$). Após dez validações cruzadas, o modelo *Artificial Neural Network* (ANN) obteve a melhor performance. O modelo *Random Forest* apresentou uma melhora de 12,0% em relação ao APACHE II score, enquanto o modelo CRT alcançou sensibilidade, especificidade, valor preditivo positivo (PPV), valor preditivo negativo (NPV) e precisão de 88,0%, 92,7%, 68,8%, 97,7% e 92,0%, respectivamente, com uma AUC-ROC de 0,90. As variáveis que mais contribuíram para os modelos foram o APACHE II score, a contagem de células sanguíneas, o tempo desde os sintomas até a admissão, a saturação de oxigênio e a contagem de linfócitos no sangue.

Ainda sobre o Covid-19, em Prakash et al. [14] tem como objetivo de identificar o grupo etário mais afetado pela doença. Entre os modelos testados, os destaques foram para o *Random Forest Regressor* e o *Random Forest Classifier*, que apresentaram um desempenho superior em comparação com outros modelos, incluindo SVM, KNN, *Decision Tree*, classificador gaussiano *Naive Bayes*, *Multiple Linear Regression*, *Logistic Regression* e *XGBoost*. Para a avaliação dos modelos, a pontuação de regressão *RSquared* (coeficiente de determinação) e a precisão foram calculadas com base na divisão do conjunto de dados em treinamento (70% dos dados) e teste (30% dos dados). Os resultados obtidos indicam que pessoas nas faixas etárias de 20 a 29, 30 a 39 e 40 a 50 anos são mais afetadas pelo Covid-19, conforme identificado pelos modelos de previsão. Além da análise dos classificadores, também foram construídos

regressores para previsão, e os resultados mostraram que tanto o *Random Forest Regressor* quanto o *Random Forest Classifier* superaram os demais modelos em termos de coeficiente de determinação.

B. Trabalhos relacionados com outras doenças

No estudo do artigo Pessoa et al. [4] é feita uma abordagem de um possível estado de choque em pacientes nas unidades de terapia intensiva (UTIs), que representa uma das principais causas de óbito. Sendo assim, uma modelagem preditiva foi realizada utilizando dados reais de pacientes, a fim de criar um modelo de predição baseado nas entradas nas UTIs. Para garantir a eficácia desse modelo, uma análise detalhada dos dados foi conduzida, identificando e removendo informações desnecessárias ou incompletas, visando uma melhor qualidade dos dados de treinamento e teste. No processo de modelagem, diversos algoritmos de classificação foram explorados, tais como *Decision Tree*, *Random Forest*, *Adaboost*, *Gradient Boosting* e *XGBoost*. Para validar o desempenho desses modelos, adotaram o método *K-Fold Cross Validation* com 5 divisões, onde os resultados foram avaliados por meio da média aritmética das métricas de *Recall*, precisão e AUC-ROC. Como resultado, observaram que os algoritmos *Gradient Boosting* e *XGBoost* apresentaram os melhores desempenhos, demonstrando o menor número de falsos negativos e, portanto, o melhor *Recall*. Notavelmente, o modelo baseado no algoritmo *XGBoost* mostrou-se especialmente eficiente, obtendo uma AUC-ROC de 1,00 e superando os demais modelos em todas as métricas de avaliação (*Recall*, precisão e AUC-ROC).

Dessa maneira, também é possível observar o mesmo estudo sobre complicações em UTIs no artigo Coelho et al. [10], mas nesse caso objetivo é prever complicações que podem surgir durante a estadia do paciente na UTI e não na sua entrada como em Pessoa et al. [4]. Os dados foram coletados de prontuários eletrônicos de três UTIs em um hospital em Belo Horizonte, abrangendo um período de 31 de julho de 2016 a 31 de dezembro de 2018, com um total de 6000 admissões e 5474 pacientes. A validação do modelo foi realizada utilizando dados de 01 de janeiro de 2019 a 30 de abril de 2019. Para a previsão de complicações, o algoritmo *LightGBM*, uma técnica de *Gradient Boosting* com árvores de decisão reforçadas, foi empregado. A otimização do modelo incluiu o ajuste de hiperparâmetros como o número de árvores no conjunto (T), a taxa de aprendizado (γ) e a profundidade máxima da árvore (θ). O desempenho dos modelos foi avaliado utilizando o método *5-Fold Cross-Validation* no desenvolvimento. Além disso, para medir a estabilidade das explicações, múltiplas iterações de amostragem com reposição foram realizadas, mostrando os dados de treinamento com reposição. Em cada iteração, uma matriz de similaridade foi criada com base nos vetores de explicação obtidos. A performance dos preditores foi avaliada usando a *Area Under the Receiver Operating Characteristic Curve*. Em todos os casos, o desempenho preditivo aumentou com a robustez e estabilidade das explicações.

A predição do risco de morte em medicina intensiva, visando melhorar a tomada de decisão terapêutica também é tema do Schmidt et al. [15]. O modelo adotado pelos pesquisadores

foi a *Neural Network Convolutional* (RNC) dividida em cinco estágios e com nove camadas ocultas. A abordagem quantitativa compara o poder preditivo do modelo proposto com o popular *Acute Physiology and Chronic Health Evaluation II* (APACHE II), amplamente utilizado em Unidades de Terapia Intensiva (UTIs). Os dados utilizados foram extraídos de registros médicos de pacientes adultos não identificados, provenientes da base de dados *Multiparameter Intelligent Monitoring in Intensive Care III* (MIMIC III). O modelo foi desenvolvido em dois conjuntos de dados, sendo o segundo com 11 variáveis adicionais em relação ao primeiro. O processo do modelo envolve camadas de convolução e subamostragem para extração de características e redução da dimensionalidade. As camadas de subamostragem utilizam *Max-Pooling* para capturar os valores mais relevantes em regiões específicas da entrada. Os resultados foram avaliados com métricas como acurácia, sensibilidade, especificidade e AUC-ROC. Comparando as versões V1 e V2 do modelo, os autores observaram desempenho semelhante na sensibilidade para identificar casos positivos de mortalidade.

O câncer de mama é o câncer mais identificado entre as mulheres e principal razão para o aumento da taxa de mortalidade entre elas. Dessa maneira, os artigos Pinheiro et al. [6] e [17] abordam a sobrevida em casos de câncer de mama, buscando identificar os parâmetros que influenciam a sobrevida dos pacientes. No primeiro, uma análise univariada foi conduzida, utilizando medidas como a mediana para variáveis quantitativas e frequência absoluta e relativa para variáveis qualitativas. Em seguida, uma análise de sobrevida foi realizada, utilizando a curva de *Kaplan-Meier* e o Teste de *Log-Rank*. Posteriormente, uma análise multivariada foi conduzida utilizando a regressão de Cox, precedida pela avaliação do pressuposto de proporcionalidade dos riscos, sendo o tamanho do efeito medido pelo *Hazard Ratio* (HR). Além disso, o algoritmo RSF (*Random Survival Forest*) foi modelado, com 70% dos dados utilizados para treinamento e 30% para teste. Os índices de concordância *C-index* e *Integrated Brier Score* (IBS) foram utilizados para avaliar o desempenho do modelo na classificação e na calibração das predições, respectivamente. Os resultados apontaram que a idade, tamanho do tumor e estadiamento foram fatores de risco importantes para a sobrevida de pacientes com câncer de mama. No segundo, a metodologia utilizada envolve a implementação de modelos de *Logistic Regression*, *Support Vector Machine* (SVM) e *K-Nearest Neighbors* (KNN) para classificação. Foram avaliados resultados de acurácia, precisão, sensibilidade, especificidade e taxa de falso positivo para comparar a eficiência de cada algoritmo. Para reduzir a dimensionalidade do conjunto de dados, foi aplicada a técnica de redução de dimensionalidade chamada *Principal Component Analysis* (PCA), que ajudou a encontrar os recursos mais relevantes com base na matriz de covariância. A seleção de características foi realizada para encontrar um subconjunto de recursos mais informativos, considerando a precisão e os erros de previsão. O conjunto de dados continha 32 atributos, e a redução de dimensionalidade foi essencial para lidar com dados multidimensionais. Dos três algoritmos aplicados pelos autores: (*Support Vector Machine-SVM*, *K-Nearest Neighbors* e *Logistic Regression*), o SVM

obteve a maior precisão, alcançando 92,7% em comparação com os outros dois algoritmos. Com base nos resultados, os pesquisadores indicaram que o SVM foi o algoritmo mais adequado para prever a ocorrência de câncer de mama com esses conjuntos de dados complexos. A análise dos resultados indicou que a integração de dados multidimensionais, juntamente com diferentes técnicas de classificação, seleção de recursos e redução de dimensionalidade, pode fornecer ferramentas promissoras para inferências neste domínio.

Procurando demonstrar o uso de algoritmos de *aprendizado de máquina* para análises preditivas em saúde, utilizando dados de idosos residentes no Município de São Paulo, o estudo de Santos et al. [18] divide os dados como em Silva [7]. No pré-processamento, os valores faltantes da variável quantitativa "índice de massa corporal" (IMC) foram imputados com a mediana do conjunto de treinamento. Para as variáveis categóricas "escolaridade" e "história de queda", que tinham valores faltantes, uma categoria adicional foi criada para representar tais observações. Para os demais preditores com valores faltantes, as observações correspondentes foram eliminadas. As variáveis quantitativas "IMC" e "idade" foram padronizadas com base na média e no desvio padrão do conjunto de treinamento. Foram utilizados cinco algoritmos na etapa de aprendizado para ajustar modelos: *Logistic Regression* com e sem penalização, *Neural Network*, *Gradient Boosted Trees* e *Random Forest*. Os hiperparâmetros dos algoritmos foram otimizados utilizando validação cruzada de 10 campos para selecionar os melhores modelos. A métrica utilizada para otimização e seleção dos modelos durante o aprendizado, como em Coelho et al. [10], foi a AUC-ROC. O modelo com maior AUC-ROC para cada algoritmo foi escolhido para a avaliação de desempenho nos dados de teste. Os cinco preditores mais importantes para os modelos selecionados apresentaram classificações similares, mas em ordens diferentes, onde "idade" apareceu como primeiro lugar em quatro preditores. Os modelos escolhidos apresentaram um desempenho satisfatório, com AUC-ROC superior a 0,70 para todos eles. As melhores performances foram obtidas pelos modelos de *Neural Network*, *Logistic Regression* com penalização de lasso (*Least Absolute Shrinkage And Selection Operator*) e *Logistic Regression* simples, sem grandes diferenças entre eles.

Em Oliveira et al. [8], o foco do estudo foi sobre a tuberculose e sua alta taxa de mortalidade. Os pesquisadores buscaram propor um modelo de *Deep Learning* para auxiliar no pós-diagnóstico da doença. A base de dados utilizada contém informações de pacientes com tuberculose, incluindo aqueles que evoluíram para óbito no estado do Amazonas entre 2007 e 2018. Para garantir um conjunto de dados balanceado, foi realizado um processo de *undersampling*, uma vez que houve menos casos de óbito em relação aos casos não fatais. A divisão dos dados em 50% para treinamento, 25% para validação e 25% para teste permitiu uma avaliação abrangente do desempenho do modelo. As métricas utilizadas para a avaliação incluíram *F1-score*, precisão, sensibilidade, especificidade e acurácia. Foram empregados dois modelos de *Deep Learning*: um modelo totalmente conectado e uma *Neural Network Convolutional* (CNN). Para cada modelo, a técnica *Gridsearching* foi utilizada para identificar as melhores configurações dos hiperparâmetros.

Os resultados mostraram que os modelos de *Deep Learning* totalmente conectados obtiveram um desempenho superior em comparação aos modelos CNN. No entanto, ambos os modelos apresentaram valores acima de 70% para quase todas as métricas, demonstrando que não há uma única configuração que alcance os melhores resultados para todas as métricas simultaneamente.

Levando em conta um outro assunto importante para a sociedade, Krishnamoorthi et al. [16] explorou quatro métodos de aprendizado de máquina com o objetivo de realizar análises preditivas relacionadas ao diabetes. No processo de pré-processamento dos dados, foram tratados e removidos dados inconsistentes, e valores ausentes foram atribuídos a alguns atributos selecionados para aumentar a precisão dos resultados. Em seguida, os dados foram normalizados e divididos em conjuntos de treinamento (90%) e teste (10%). Os modelos aplicados foram *Logistic Regression* (LR), *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM) e *Random Forest*. Para aprimorar a performance dos algoritmos, uma validação cruzada e ajustes de hiperparâmetros (variáveis que controlam o próprio processo de treinamento) foram realizados utilizando técnicas de *Gridsearching* e *Random Search*. Dessa forma, métricas como sensibilidade, especificidade e precisão foram calculadas e como em Assaf et al. [13], a curva ROC foi empregada para avaliar o desempenho dos algoritmos, destacando que o modelo de *Logistic Regression* obteve o melhor desempenho com um valor de ROC de 86%. Como resultado do estudo, os autores observaram uma correlação entre os níveis de glicose e o Índice de Massa Corporal (IMC) com a ocorrência de diabetes.

O resumo abrangente dos artigos analisados durante a revisão bibliográfica pode ser encontrado na tabela I. Cada linha nela representa um artigo específico e inclui informações essenciais como as referências bibliográficas usadas no estudo, os tipos de dados usados nas análises, a técnica de aprendizado escolhida para o projeto, os resultados alcançados e também uma abordagem do problema em específico. A tabela I desempenha um papel importante nos estudos revisados ao servir como um guia referencial que apresenta uma visão consolidada das diferentes abordagens e descobertas através do emprego das técnicas de aprendizado de máquina.

IV. EXPERIMENTO

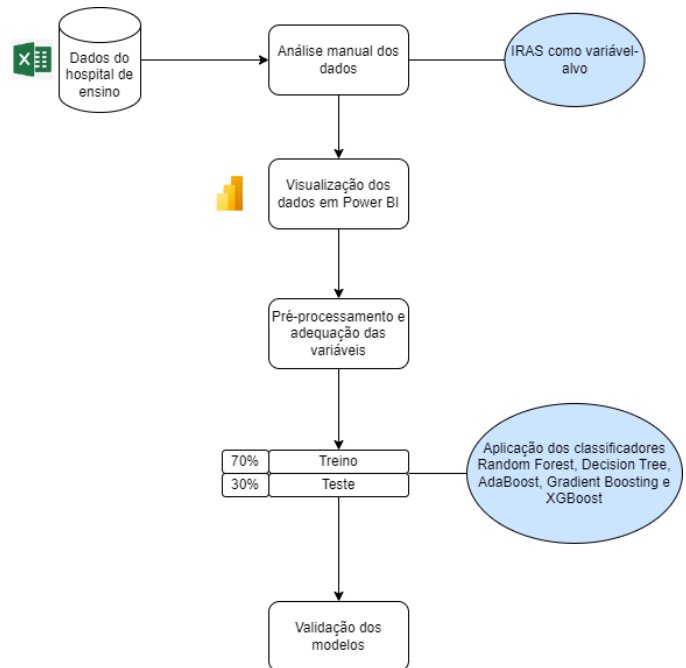
A. Desenho do Experimento

Na figura 2, é possível visualizar o experimento realizado desde a primeira etapa de aquisição de dados até o estágio final de validação do modelo. Este diagrama adota todas as etapas essenciais do processo, incluindo a visualização dos dados iniciais, o pré-processamento para limpeza e preparação, a fase de treinamento do modelo e a verificação do mesmo no final. Esta figura oferece uma visão geral do processo, ilustrando o modo como as tarefas encadeiam-se umas nas outras.

B. Dados do Hospital

Este estudo utilizou-se de dados secundários anonimizados de base de dados disponíveis no serviço de controle de Infecção Relacionada à Assistência à Saúde, de um Hospital das Clínicas

Fig. 2. Desenho do experimento.



de Goiânia, Goiás. Os dados utilizados são de todos os casos notificados de infecções. Os dados se referem ao período de janeiro a agosto de 2021, sendo 792 notificações. Ao admitir um paciente no hospital, é iniciada a anotação de suas informações, que incluem dados como microrganismos, antimicrobianos e outros tratamentos durante a permanência do paciente. O presente estudo foi realizado em parceria com o Núcleo de Estudo e Pesquisa de Enfermagem em Prevenção e Controle de Infecções Relacionadas à Assistência à Saúde (NEPIH) por meio do projeto de pesquisa intitulado de “Indicadores epidemiológicos das infecções relacionadas aos cuidados em saúde (IRAS) no município de Goiânia Go”, aprovado pelo Comitê de Ética em Pesquisa (protocolo de aprovação nº 1269485/2015).

De fato, a clareza na seleção das colunas é importante para orientar nas análises subsequentes. Dessa maneira, essas colunas escolhidas servirão como base para os experimentos realizados. A tabela II fornece uma visão detalhada das variáveis totais utilizadas. Isso contribui para uma compreensão mais orientada dos elementos empregados nas análises e conclusões apresentadas neste estudo.

C. Análise manual dos dados

Para ressaltar a importância, é preciso compreender os dados hospitalares antes da aplicação de quaisquer técnicas mais avançadas. Assim, foi feito um estudo manual de cada coluna e o significado de seus valores. Vale ressaltar que nessa análise foi empregada a ferramenta *Power BI*, com destaque para a importância no processo de visualização e compreensão dos dados relacionados à área da saúde.

D. Preparação e adequação dos dados

Durante o processo de pré-processamento dos dados para este estudo, foram implementadas várias etapas importantes para

Tabela II. Variáveis disponíveis para o experimento.

Características	Definição
IDADE	O tempo de vida decorrido desde o nascimento até uma determinada data tomada como referência.
TOPOGRAFIA	Descrição sistemática de sítios geográficos específicos, em relação aos aspectos físicos que se acredita influenciarem sobre a saúde e a doença.
SEXO	Refere às características biológicas e físicas que diferenciam os seres humanos como machos ou fêmeas.
DIAGNÓSTICO	É o processo de identificação e determinação da natureza de uma doença ou condição médica com base em sinais, sintomas, exames clínicos, laboratoriais e outras informações relevantes.
MICROORGANISMO	É um organismo microscópico que pode ser encontrado em uma variedade de ambientes, incluindo solo, água, ar e organismos vivos.
ÓBITO	Termo usado para descrever a morte de um indivíduo.
CVC	Cateter Venoso Central, no qual é colocado de um cateter com a sua extremidade posicionada na veia cava superior ou no átrio direito. Para facilitar, assim, o acesso vascular em pacientes que necessitam de terapia intravenosa prolongada ou acesso venoso para administração de medicamentos, nutrição parenteral, hemodiálise, monitoramento hemodinâmico, entre outros fins médicos.
PERMANÊNCIA CVC (DIAS)	Total de dias aplicando o Cateter Venoso Central.
VM	Ventilação Mecânica que consiste no uso de uma máquina para auxiliar a entrada e a saída de ar nos pulmões.
PERMANÊNCIA VM (DIAS)	Total de dias aplicando a Ventilação Mecânica.
SVD	Sonda vesical de Demora que consiste em manter um cateter para drenagem contínua da urina do interior da bexiga.
PERMANÊNCIA SVD (DIAS)	Total de dias aplicando a Sonda vesical de Demora.
SNE	Sonda Nasoenteral com finalidade de alimentar e hidratar o paciente pelo nariz.
PERMANÊNCIA SNE (DIAS)	Total de dias aplicando a Sonda Nasoenteral.
NPT	Administrada através de um cateter colocado em uma veia, ela é usada quando o paciente não consegue obter todos os nutrientes necessários pela boca ou pela alimentação por tubo.
PERMANÊNCIA NPT (DIAS)	Total de dias aplicando a Nutrição Parenteral Total.
DRENOS	Dispositivos de formas variadas, utilizados para saída de ar e líquidos que podem se acumular em uma cavidade.
PERMANÊNCIA DRENOS (DIAS)	Total de dias aplicando os drenos.
ANTIMICROBIANO 1, 2, 3, ..., 16.	Cada número representa um antimicrobiano diferente, aqui foi colocado desta forma para melhorar a visualização. Um antimicrobiano é um medicamento utilizado para tratar as infecções, em particular, as de origem bacteriana e o fato de serem muitos tipos se dá devido à diversidade de agentes infecciosos.
DIAS TOTAIS ANTIMICROBIANO 1, 2, 3, ..., 16.	Total de dias em que o antimicrobiano em específico foi aplicado.
IRAS (variável-alvo)	Infecção Relacionada à Assistência a Saúde é qualquer infecção adquirida após a admissão do paciente no hospital.

garantir a consistência e qualidade das informações utilizadas. No começo, escolheu-se pela padronização e adequação dos valores incompletos. Tal ação foi indispensável para prevenir disparidades resultantes da formatação original dos registros. Adicionalmente, técnicas foram aplicadas para eliminar pontuação inadequada, a fim de assegurar que os dados estejam uniformes e evitar qualquer confusão na interpretação. A estratégia adotada para preencher lacunas incompletas nos dados foi atribuir o valor zero. Essa abordagem foi utilizada com o intuito de informar que não foi identificada uma determinada característica no paciente, assegurando assim consistência na interpretação dos dados ausentes.

O passo mais importante foi converter os valores textuais em representações numéricas, pois isso viabilizou a aplicação

eficiente dos algoritmos de aprendizado de máquina. Foi adotada a técnica do *bag of words* para extrair as informações de colunas com variáveis com subconjuntos parecidos. O método foi aplicado nas variáveis da tabela III. Este passo foi essencial nos resultados dos modelos de aprendizado de máquina conseguirem identificar similaridades entre os valores dessas colunas.

Para a transformação das palavras extraídas do *bag of words* em números binários, utilizou-se o *MultiLabelBinarizer*. Essa etapa possibilita o treinamento dos modelos, garantindo que pudessem trabalhar de maneira eficiente com as características extraídas. Estabelecendo uma base sólida para as análises subsequentes, a abordagem adotada durante o pré-processamento dos dados assegura a consistência, confiabilidade e adaptabilidade

Tabela III. Colunas que receberam a técnica do *bag of words*.

Colunas com <i>bag of words</i>		
MICROORGANISMO	DIAGNÓSTICO	TOPOGRAFIA
ANTIMICROBIANO 1	ANTIMICROBIANO 2	ANTIMICROBIANO 3
ANTIMICROBIANO 4	ANTIMICROBIANO 5	ANTIMICROBIANO 6
ANTIMICROBIANO 7	ANTIMICROBIANO 8	ANTIMICROBIANO 9
ANTIMICROBIANO 10	ANTIMICROBIANO 11	ANTIMICROBIANO 12
ANTIMICROBIANO 13	ANTIMICROBIANO 14	ANTIMICROBIANO 15
ANTIMICROBIANO 16		

dos dados ao contexto do estudo.

Antes de realizar o processamento dos dados, uma análise abrangente da qualidade foi feita para identificar valores faltantes. Em encontros com a Dra. Silvana, foram exploradas as razões sobre a ausência de informações em determinadas colunas. Verificou-se que a falta de dados em algumas características específicas dos pacientes sugeria a não ocorrência dessas características, sendo então substituídos por "0" para refletir essa ausência. Por outro lado, em alguns casos, a ausência de dados era inevitável, demandando a aplicação de técnicas específicas para lidar com esse cenário.

Diante do desafio de lidar com um desbalanceamento significativo na coluna Infecção Relacionada à Assistência à Saúde (IRAS), foi implementado o *RandomOverSampler* durante o pré-processamento dos dados. Esse método foi escolhido devido à diferença nas proporções das classes 0 (734 pacientes que não adquiriram IRAS no hospital) e 1 (57 pacientes que contraíram alguma IRAS), visando mitigar os impactos negativos desse desbalanceamento na performance dos modelos de aprendizado de máquina. Assim, a utilização do *RandomOverSampler* gerou um conjunto de dados balanceado, garantindo que ambas as classes fossem representadas de maneira mais equitativa, favorecendo a capacidade dos modelos de aprenderem efetivamente as características associadas à classe minoritária.

É importante levantar o ponto de que a comparação entre diferentes técnicas de amostragem revelou desempenhos diferentes dos modelos. Ao se empregar um método de *undersampling*, observou-se uma queda na acurácia e precisão, além de um aumento da taxa de falso negativo e falso positivo da matriz de confusão. Isso sugere que a remoção de instâncias da classe majoritária comprometeu a capacidade dos modelos em generalizar padrões presentes nos dados. Assim sendo, o *RandomOverSampler* não apenas contribuiu para balancear as classes, mas também mitigou os desafios associados ao desbalanceamento, promovendo uma performance mais robusta e confiável na previsão de IRAS.

E. Seleção de características

Após concluir os estágios iniciais de preparação dos dados, uma etapa fundamental foi o estudo das características mais importantes para a predição. A figura 3 apresenta um resumo das variáveis significativas para a predição de IRAS. As variáveis "MICROORGANISMO", "TOPOGRAFIA" e "IDADE" foram elencadas como as 3 mais influentes, sugerindo que desempenham papéis consideráveis na explicação das variações nos dados. A presença de microrganismos demonstrou ser

significativa por conta da quantidade de valores correlacionados, indicando uma forte influência com o resultado. A topografia, por sua vez, foi consistentemente selecionada nas divisões das árvores de decisão, sugerindo sua importância nos processos modelados. A idade, como terceira variável destacada, apontou para padrões específicos relacionados ao estágio de vida que influenciam as respostas previstas. Isto foi determinado utilizando métodos de *Feature Importance* do *Random Forest*, que avalia a contribuição relativa de cada variável para a precisão do modelo. A análise da importância das características permite entender quais atributos têm maior peso na decisão do modelo, oferecendo percepções valiosas sobre os fatores que mais impactam na ocorrência de IRAS.

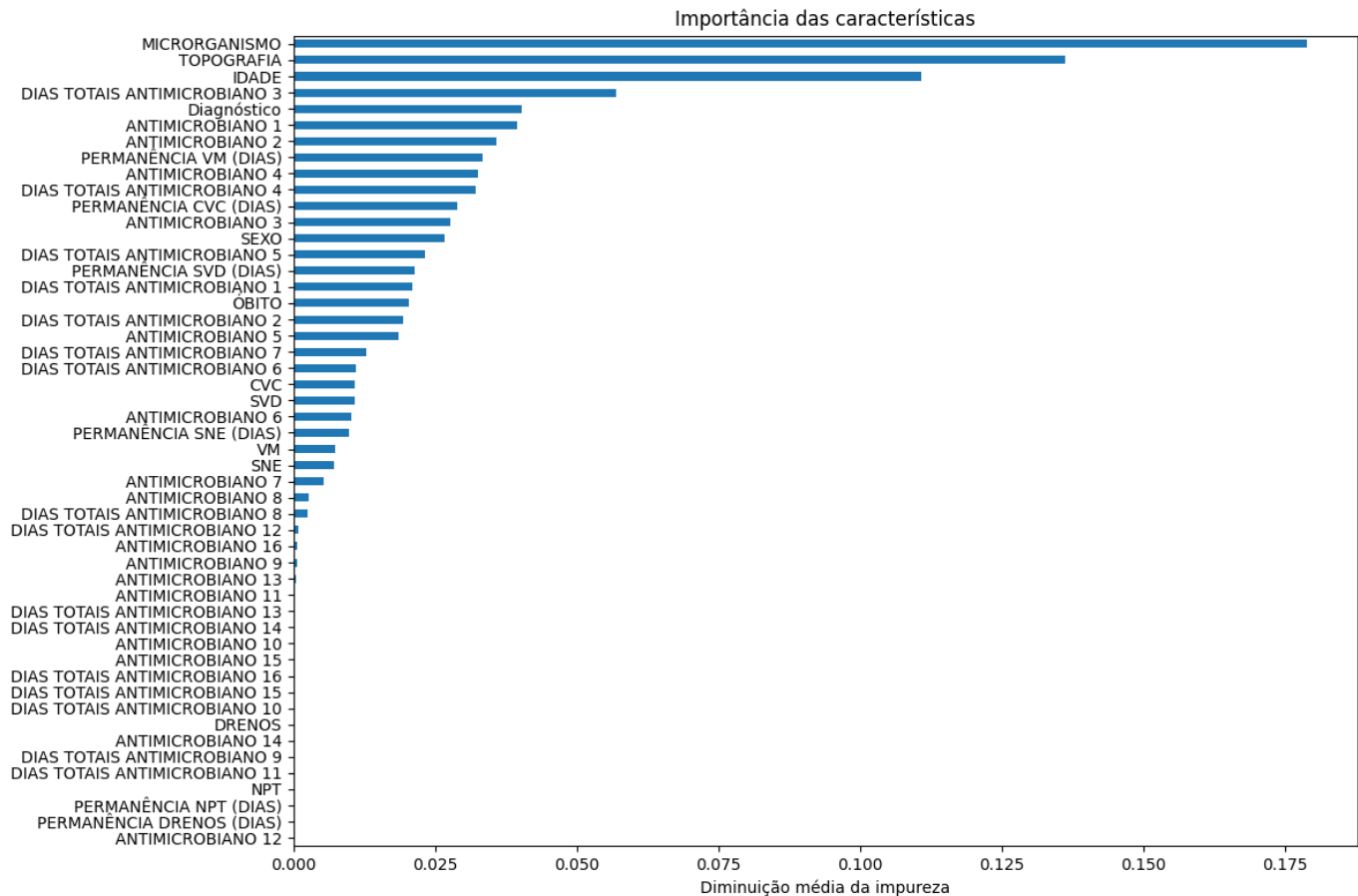
Dessa maneira, a análise preditiva de dados na UTI se apresenta como uma ferramenta importante para a equipe de enfermagem ao possibilitar a antecipação de complicações por meio da identificação precoce do estado de saúde dos pacientes. Esta capacidade de resposta ágil pode evitar situações críticas, melhorando os resultados clínicos. Além disso, a compreensão profunda das variáveis contribui para a otimização de recursos, promovendo eficiência na operação, gestão de medicamentos e equipamentos. Essa abordagem visa aprimorar a qualidade do atendimento, reduzindo desperdícios e fortalecendo a segurança do paciente.

Dentre os dados, a coluna sobre IRAS foi selecionada como variável-alvo. Essa decisão estratégica reflete o foco do estudo, em prever precisamente a condição de infecção relacionada à saúde usando vários aspectos extraídos do conjunto de informações. A etapa de mineração das características resultou na identificação de 50 variáveis preditoras como mostrado na tabela II, no qual a variável em azul é a variável-alvo e as em branco são as preditoras. Antes de extrair as características, foi utilizada uma técnica de pré-processamento que converte valores categóricos em numéricos. Muitos algoritmos de aprendizado de máquina precisam que os dados de entrada sejam numéricos, e o *LabelEncoder* é uma ferramenta útil para mapear cada categoria para um valor numérico único. A figura 3 mostra as características mais importantes utilizadas no treinamento do modelo.

F. Treinamento e validação do modelo

Durante as fases de treinamento e teste dos dados, foi essencial adotar uma abordagem estratégica, onde 70% das amostras foram usadas para o treinamento, enquanto os 30% restantes foram reservados para os testes. Esse mesmo esquema também foi utilizado nos artigos Pinheiro et al. [6], Silva [7] e Prakash et al. [14]. Essa divisão estratégica garante uma

Fig. 3. Características mais importantes apresentadas no dados.



avaliação sólida do desempenho do modelo e também permite uma análise completa de sua capacidade de generalização para novos dados.

É relevante enfatizar a conversão em números binários de determinadas colunas das características preditoras. Esse processo é necessário para garantir uma boa performance no treinamento dos modelos de aprendizado de máquina. Para resolvermos o problema do desequilíbrio na variável-alvo IRAS, foi preciso utilizar a técnica *RandomOversampler*. Dessa forma foi possível obter um conjunto de dados balanceados onde é visto exatamente as classes iguais: 734 amostras da classe 1 (adquiriu IRAS) e mantendo a classe 0 (não adquiriu IRAS) no total de 734 amostras também. Como visto no estudo de Dixit [19], para evitar viés decorrente da diferença entre as classes, é importante buscar esse equilíbrio nas amostras e criar uma representação mais equilibrada dos resultados

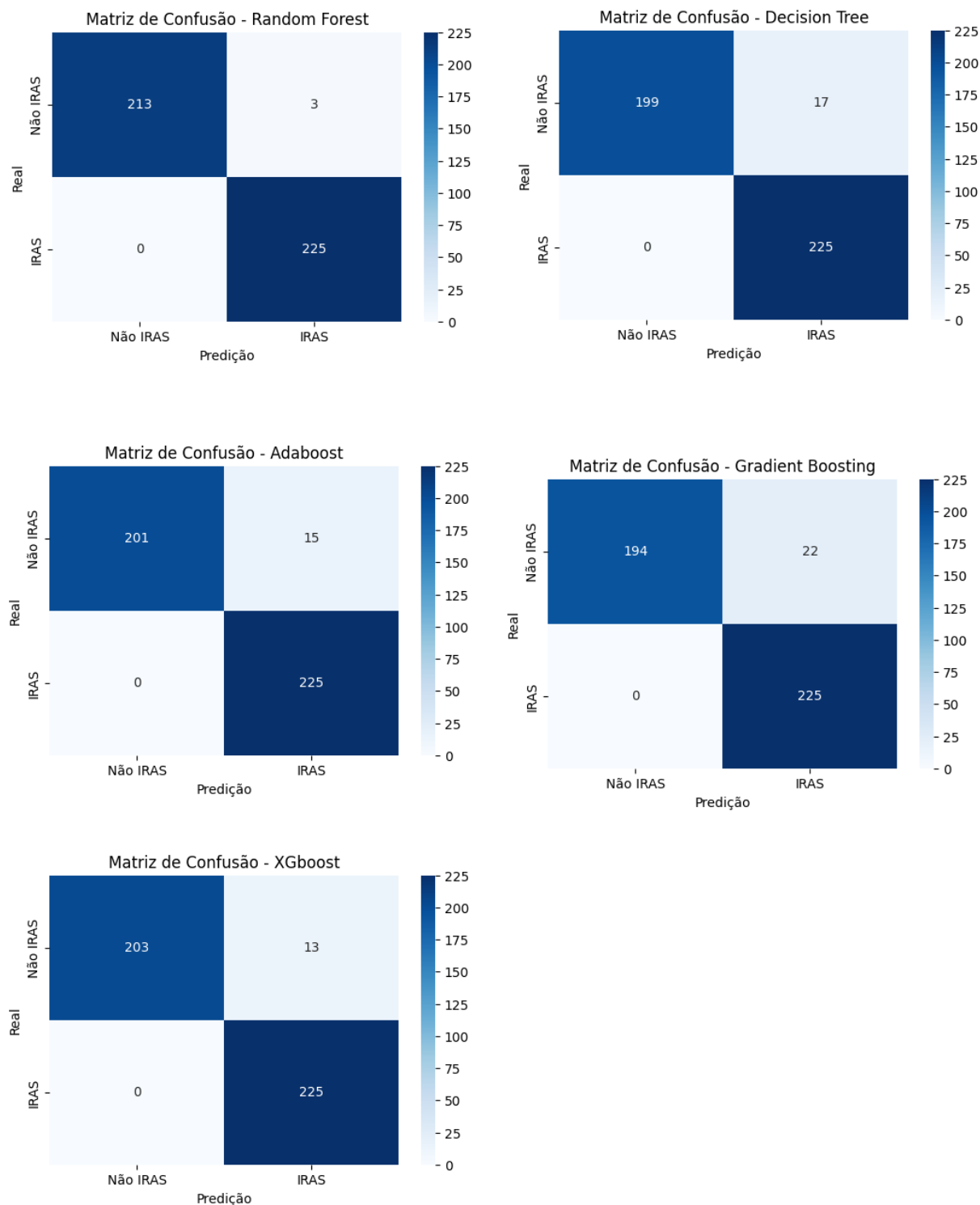
1) *Aplicação dos modelos de aprendizagem*: Inicialmente, empregou-se a técnica de *Random Forest* (RF) para apresentar as características mais relevantes antes do treinamento dos modelos subsequentes. A sequência de algoritmos *Random Forest* (RF), *Adaboost* (AB), *XGBoost* (XG), *Gradient Boosting* (GB) e *Decision Tree* (DT) foi aplicada em um estágio posterior. E utilizando a técnica *K-Fold Cross Validation*, procurou-se otimizar os modelos para melhorar o desempenho em termos de resposta. Ao lidarmos com modelos de aprendizado de máquina, podemos observar que um aspecto importante da RF

é sua habilidade para mitigação do sobreajuste e melhora na acurácia das previsões. Já no caso da DT temos como principal característica uma abordagem compreensível construída através de decisões hierárquicas. Métodos de *ensemble* como AB e GB visam melhorar o desempenho combinando modelos mais fracos. Por outro lado, o XG introduz inovações como regularização e otimização eficiente para elevar ainda mais a capacidade preditiva. Para explorar diferentes abordagens e identificar a mais adequada para esse estudo, foi feita a escolha dessa variedade de modelos, que possibilita uma análise mais abrangente sobre as capacidades preditivas em relação às IRAS apresentadas.

2) *Validação dos modelos*: A fim de avaliar cada modelo de aprendizado de máquina, foi incluída na análise a técnica de *K-Fold Cross Validation*. Foi definido o parâmetro de $k = 5$, que representa a quantidade na qual a base de dados foi dividida, para uma validação cruzada em 5 vezes. O desempenho do modelo pode ser compreendido de maneira mais completa através dessa abordagem, que minimiza os efeitos da aleatoriedade na seleção dos conjuntos de treinamento e validação.

Assim, métricas como a matriz de confusão, acurácia, precisão, *F1-score*, *Recall* e a curva AUC-ROC foram calculadas. Essa abordagem proporcionou uma análise abrangente dos modelos na previsão de IRAS, preparando o terreno para uma compreensão mais detalhada dos resultados obtidos.

Fig. 4. Matriz de confusão para cada modelo.



V. RESULTADOS E DISCUSSÃO

A capacidade de realizar análises aprofundadas e fazer referências apropriadas não só desempenha um papel valioso na produção científica, mas também se revela essencial para a evolução constante do entendimento em múltiplas esferas do conhecimento. A tabela IV detalha o desempenho dos

modelos antes do balanceamento de dados, enquanto a tabela V apresenta o desempenho após a realização do balanceamento. A análise comparativa entre esses conjuntos de resultados oferece pontos valiosos sobre o impacto do balanceamento de dados nas métricas de desempenho dos modelos.

A interpretação desses resultados ajuda a compreender tanto

Tabela IV. Desempenho dos modelos antes do balanceamento de dados.

Modelo	Acurácia	Precisão	Recall	F1-score
<i>Random Forest</i>	0.9571	1.0000	0.2153	0.3533
<i>Decision Tree</i>	0.9361	0.4202	0.3846	0.3915
<i>Adaboost</i>	0.9386	0.4405	0.4461	0.4396
<i>Gradient Boosting</i>	0.9537	0.6299	0.4153	0.4967
<i>XGBoost</i>	0.9386	0.4492	0.5076	0.4758

Tabela V. Desempenho dos modelos após a realização do balanceamento de dados.

Modelo	Acurácia	Precisão	Recall	F1-score
<i>Random Forest</i>	0.9913	0.9834	1.0000	0.9916
<i>Decision Tree</i>	0.9673	0.9399	1.0000	0.9690
<i>Adaboost</i>	0.9637	0.9424	0.9893	0.9652
<i>Gradient Boosting</i>	0.9537	0.9184	0.9982	0.9565
<i>XGBoost</i>	0.9750	0.9534	1.0000	0.9761

as áreas em que o modelo está acertando como aquelas em que está cometendo erros. Acurácia, precisão, *Recall* e *F1-score* são algumas das métricas obtidas a partir da matriz de confusão que nos ajudam a entender o desempenho do modelo. A precisão indica a proporção de instâncias corretamente classificadas em relação a todas as instâncias previstas como positivas. A capacidade de identificar corretamente todas as instâncias positivas é medida através do *Recall* ou sensibilidade. A métrica do *F1-score* combina a precisão e o *Recall* em uma média harmônica. Assim, é essencial analisar a matriz de confusão para compreender as implicações práticas das previsões do modelo e garantir que ele esteja adequado aos objetivos do problema de classificação em questão.

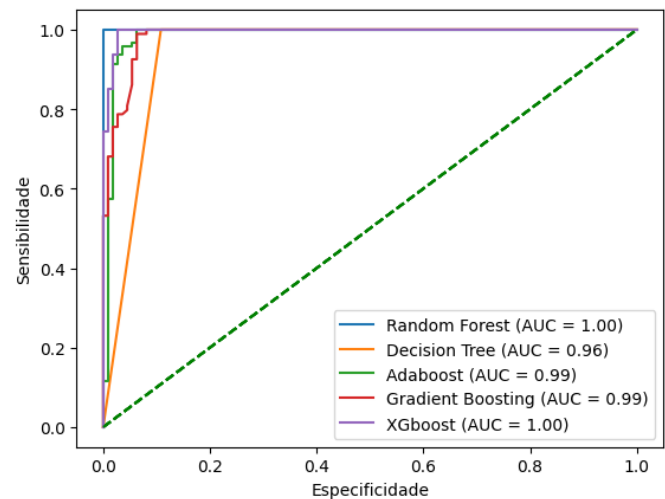
Após o treinamento, constatou-se que o modelo *Random Forest* se destacou como sendo inicialmente a abordagem mais eficiente para realizar a tarefa desejada antes da utilização da técnica de balanceamento *RandomOverSampler*. A capacidade do *Random Forest* em tratar a distribuição desbalanceada do conjunto de dados e produzir bons resultados foi demonstrada pela observação inicial, onde o algoritmo obteve uma acurácia de 95%, precisão de 100%, *Recall* de 21% e *F1-score* de 35%. Os resultados mostraram uma melhoria notável após a introdução do *RandomOverSampler* como solução para o problema de desbalanceamento. Mais uma vez, o modelo *Random Forest* mostrou-se mais eficiente após a aplicação do *oversampling* com acurácia de 99%, precisão de 98%, *Recall* de 100% e *F1-score* de 99%. Isso demonstra o papel importante dessa abordagem na melhoria das previsões.

Ao modelar matrizes de confusão, é possível obter *insights* sobre a qualidade de cada modelo. Na matriz de confusão da *Random Forest*, a distribuição foi de 3 falsos positivos, o que indica que o modelo classificou incorretamente três pacientes como portadores de IRAS que, na verdade, não eram. E assim, a *Decision Tree* apresentou 17 falsos positivos, seguida pelo *Adaboost* com 15, o *Gradient Boosting* com 22 e o *XGBoost* com 13. Em todas essas situações, os modelos classificaram pacientes como portadores de IRAS de forma inadequada. O mais significativo, foi que nenhum dos modelos apresentou falsos negativos, sugerindo que todos os pacientes com IRAS foram corretamente identificados pelos modelos,

como apresentado na figura 4.

Neste contexto específico do trabalho realizado, o desempenho superior do *Random Forest* pode ser atribuído a várias características desse modelo que mostraram-se benéficas para o conjunto de dados em questão. A estrutura baseada em múltiplas árvores proporcionou uma maior robustez e resistência a *overfitting*, resultando em um desempenho geral melhor. Também foram fatores determinantes a capacidade do algoritmo em lidar naturalmente com conjuntos de dados com muitas características e sua versatilidade em adaptar-se a diferentes distribuições de dados.

Fig. 5. Curva Receiver Operating Characteristic dos modelos.



Como avaliação do desempenho dos modelos, é possível observar na curva ROC como a taxa de verdadeiros positivos se relaciona com a taxa de falsos positivos. Já a AUC avalia, numericamente, o poder discriminativo do modelo. Um modelo considerado excelente tem seu valor na AUC-ROC próximo de 1, já valores próximos de 0,5 sugerem um desempenho similar ao de um classificador aleatório. Logo, os resultados obtidos, como apresentando na figura 5, mostram que tanto *Random Forest* quanto *XGBoost* atingiram uma AUC de 1.0, o que mostra muito bem sua capacidade de discriminação

entre classes. A Decision Tree, por outro lado, teve uma AUC de 0,96, no qual ainda é considerado bastante decente, mas ligeiramente mais próximo da diagonal 0,5. Sendo assim, a AUC-ROC se mostra como uma ferramenta valiosa na avaliação da capacidade do modelo de diferenciar entre as classes, assim tomando decisões importantes sobre sua efetividade em cenários práticos.

VI. CONCLUSÃO

A infecção em UTIs é um problema importante que conduz pacientes a óbito. Este ambiente apresenta crescente complexidade e contínuo esforço para a melhoria da qualidade no cuidado da saúde. Antecipar a ocorrência dessa contaminação é uma ação importante para tal melhoria. Os dados de UTI são volumosos, complexos e incompletos, tornando sua análise difícil em tempo hábil para a tomada de decisão sobre ações preventivas, além do cansaço humano que realiza o cuidado. A computação é uma ferramenta importante nessa análise.

O objetivo deste trabalho é explorar algoritmos de aprendizado de máquina para a antecipação do desenvolvimento de infecção em pacientes de UTI para apoiar os profissionais da saúde na tomada de decisões. Os algoritmos abordados foram escolhidos por suas capacidades em lidar com conjuntos de dados complexos e não lineares. O que apresentou melhor desempenho foi o *Random Forest*. A característica de trabalho que combina várias árvores de decisão para obter resultados mais precisos deste algoritmo foi determinante para lidar com as características de variáveis distintas, faltantes e complexas dos dados.

Durante a análise dos dados, produziu-se um ranqueamento em termos de maior impacto à realização da previsão de desenvolvimento de infecção. As características “MICRORGANISMO”, “TOPOGRAFIA” e “IDADE” foram as 3 mais importantes definidas pelo algoritmo de *Feature Importance* do *Random Forest*. Essa observação, visa aprimorar a tomada de decisões clínicas, possibilitando a implementação de medidas preventivas eficazes e, assim, contribuindo para a redução significativa de riscos e complicações associadas às IRAS, promovendo, assim, a segurança e melhoria na qualidade global do atendimento aos pacientes.

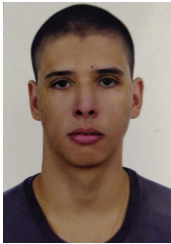
REFERÊNCIAS

- [1] P. J. Z. Teixeira, F. T. Hertz, D. B. Cruz, F. Caraver, R. C. Hallal, and J. d. S. Moreira, “Pneumonia associada à ventilação mecânica: impacto da multirresistência bacteriana na morbidade e mortalidade,” *Jornal Brasileiro de Pneumologia*, vol. 30, pp. 540–548, 2004.
- [2] Understanding Random Forest/. [Online]. Available: <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>
- [3] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random forests and decision trees,” *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, p. 272, 2012.
- [4] S. M. B. Pessoa, B. S. d. S. Oliveira, W. G. d. Santos, A. N. M. Oliveira, M. S. Camargo, D. L. A. B. d. Matos, M. M. L. Silva, C. C. d. Q. Medeiros, C. S. d. S. Coelho,

- J. d. S. Andrade Neto *et al.*, “Predição de choque séptico e hipovolêmico em pacientes de unidade de terapia intensiva com o uso de machine learning,” *Revista Brasileira de Terapia Intensiva*, vol. 34, pp. 477–483, 2023.
- [5] D. L. Freire, R. de Oliveira, J. Carmelo Filho *et al.*, “Machine learning applied in sars-cov-2 covid 19 screening using clinical analysis parameters,” *IEEE Latin Am. Trans.*, vol. 100, no. 1, 2020.
- [6] T. S. Pinheiro, E. Yahata, P. D. dos Santos, F. S. de Oliveira, A. K. Takahata, R. Suyama, H. Tanaka, T. R. Oliveira, A. P. Romani, and P. W. Simoes, “Machine learning e análise multivariada aplicados à sobrevida do câncer mama,” *Journal of Health Informatics*, vol. 14, 2022.
- [7] A. O. d. Silva, “Uso de machine learning para previsão da evolução de casos de srag incluindo casos de covid-19 considerando variáveis clínicas e demográficas,” B.S. thesis, Universidade Tecnológica Federal do Paraná, 2021.
- [8] T. T. Oliveira, E. Rocha, T. Lynn, V. Sampaio, and P. T. Endo, “Análise de modelos de deep learning para auxílio no pós-diagnóstico de tuberculose,” *Boletim do Tempo Presente*, vol. 8, no. 02.
- [9] L. S. Vianna and J. de Amorim Busana, “Predição de casos de covid-19 nos municípios de santa catarina utilizando redes neurais recorrentes,” *Saúde Coletiva (Barueri)*, vol. 11, no. 67, pp. 6851–6862, 2021.
- [10] T. A. Coelho *et al.*, “Regularização de modelos para predição precoce: um estudo na predição de complicações na uti,” 2022.
- [11] A. N. Pantaleão, C. S. Filipin, L. B. Costa, L. C. Teixeira, R. A. Avendanha, T. L. Fernandes, J. Gaspar, and Z. S. N. Reis, “Aplicação de machine learning na predição do prognóstico da covid-19 em pacientes hospitalizados: uma revisão sistemática,” 2022.
- [12] I. M. Eisenlohr, “Uso de modelos de aprendizado de máquina para análise e predição de recuperação de pacientes de covid-19 no município de vitória-es,” 2022.
- [13] D. Assaf, Y. Gutman, Y. Neuman, G. Segal, S. Amit, S. Gefen-Halevi, N. Shilo, A. Epstein, R. Mor-Cohen, A. Biber *et al.*, “Utilization of machine-learning models to accurately predict the risk for critical covid-19,” *Internal and emergency medicine*, vol. 15, pp. 1435–1443, 2020.
- [14] K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar, and Y. Pawan, “Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms,” *International Journal*, vol. 8, no. 5, pp. 2199–2204, 2020.
- [15] D. Schmidt, D. B. da Silva, C. A. da Costa, and R. da Rosa Righi, “Um modelo de predição de mortalidade em unidades de terapia intensiva baseado em deep learning,” in *Anais do XVIII Simpósio Brasileiro de Computação Aplicada à Saúde*. SBC, 2018.
- [16] R. Krishnamoorthi, S. Joshi, H. Z. Almarzouki, P. K. Shukla, A. Rizwan, C. Kalpana, B. Tiwari *et al.*, “A novel diabetes healthcare disease prediction framework using machine learning techniques,” *Journal of Healthcare Engineering*, vol. 2022, 2022.
- [17] C. Shravya, K. Pravalika, and S. Subhani, “Prediction of breast cancer using supervised machine learning tech-

niques,” *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 6, pp. 1106–1110, 2019.

- [18] H. G. d. Santos, C. F. d. Nascimento, R. Izbicki, Y. A. d. O. Duarte, P. Chiavegatto Filho, and A. Dias, “Machine learning para análises preditivas em saúde: exemplo de aplicação para prever óbito em idosos de são paulo, brasil,” *Cadernos de Saúde Pública*, vol. 35, p. e00050818, 2019.
- [19] R. R. Dixit, “Predicting fetal health using cardiocograms: A machine learning approach,” *Journal of Advanced Analytics in Healthcare Management*, vol. 6, no. 1, pp. 43–57, 2022.



Vitor Pires Silva e Souza é estudante de Engenharia de Computação pela Universidade Federal de Goiás. Possui foco na fascinante interseção entre Engenharia e Ciência de Dados, com ênfase na aplicação de técnicas de Machine Learning para análise de dados.



Deborah S. A. Fernandes trabalha no Instituto de Informática da Universidade Federal de Goiás – Campus Samambaia, Goiânia, Goiás. Além disso, é Cientista da Computação (Pontifícia Universidade Católica de Goiás), mestre em Engenharia Elétrica com ênfase em Visão Computacional (Escola de Engenharia – Universidade de Brasília/DF) e doutora em Engenharia de Sistemas Eletrônicos e Automação (Escola de Engenharia -Universidade de Brasília/DF) com ênfase em análise de dados de redes sociais em sistemas de apoio à tomada de decisão.