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

Vitor Pires Silva e Souza[1], Deborah Silva Alves Fernandes[1], Silvana L. V. dos Santos[2], Márcio Giovane Cunha Fernandes[3]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brasil

[2]Faculdade de Enfermagem e Nutrição – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brasil

[3]Sistema de Informação – Universidade Estadual de Goiás (UEG)
Goiânia – GO – Brasil

vitorpirs@discente.ufg.br, deborah.fernandes@ufg.br,

silvanalvsantos@gmail.com, marcio.giovane@ueg.br

Abstract. *This article explores the use of artificial intelligence techniques to predict Healthcare-Associated Infections (HAIs). The study, based on data from a reference teaching hospital collected between January and August 2021, investigates which machine learning algorithms are most effective in predicting HAIs. Classification algorithms were used, such as Random Forest, Decision Tree, Gradient Boosting, Adaboost and XGboost. The metric of the area under the ROC curve (Receiver Operating Characteristic) and StratifiedKFold were used to measure the performance of the models. The results for Random Forest, Decision Tree, Adaboost, Gradient Boosting and XGboost were 0.91; 0.78; 0.81; 0.92; and 0.87, respectively. With this information, the study contributes to the development of strategies that reduce the risks associated with hospital infections.*

Resumo. *Este artigo explora o uso de técnicas de inteligência artificial para prever Infecções Relacionadas à Assistência à Saúde (IRAS). O estudo, baseado em dados de um Hospital Escola de referência coletados entre janeiro e agosto de 2021, investiga quais algoritmos de aprendizado de máquina são mais eficazes para prever IRAS. Foram utilizados algoritmos de classificação, como Random Forest, Decision Tree, Gradient Boosting, Adaboost e XGboost. A métrica da área sob a curva ROC (Receiver Operating Characteristic) e StratifiedKFold foram utilizados para medir o desempenho dos modelos. Os resultados para Random Forest, Decision Tree, Adaboost, Gradient Boosting e XGboost foram 0,91; 0,78; 0,81; 0,92; e 0,87, respectivamente. Com essas informações, o estudo contribui para o desenvolvimento de estratégias que reduzem os riscos associados a infecções hospitalares.*

1. Introdução

As Infecções Relacionadas à Assistência à Saúde (IRAS) representam um problema significativo em ambientes hospitalares, impactando a qualidade do atendimento médico.

Essas infecções podem ocorrer durante a prestação de cuidados médicos e estão associadas a diversos fatores. Com o avanço da inteligência artificial, surgiram novas oportunidades para melhorar a capacidade de previsão e, consequentemente, a prevenção de IRAS. Ferramentas preditivas que utilizam algoritmos de aprendizado de máquina fornecem informações úteis para profissionais de saúde tomarem decisões mais informadas. Utilizando algoritmos como *Random Forest*, *Decision Tree*, *Gradient Boosting*, *Adaboost* e *XGboost*, foi analisado dados de pacientes de um Hospital Escola no período de janeiro a agosto de 2021. A avaliação dos modelos foi realizada usando a área sob a curva ROC, além de validação cruzada por meio da técnica *StratifiedKFold*. A pesquisa foi realizada em colaboração com um Núcleo de Estudos e Pesquisas em Enfermagem para Prevenção e Controle de IRAS, contando com a aprovação do Comitê de Ética em Pesquisa.

2. Referencial Teórico

Os modelos de aprendizado de máquina mencionados incluem a *Decision Tree*, que divide os dados em subconjuntos cada vez mais puros com base em características específicas. *Random Forest* constrói várias árvores de decisão e combina suas previsões para evitar o *overfitting* (onde um modelo se ajusta muito bem aos dados de treinamento, mas não consegue generalizar bem para dados não vistos), como visto em [Ali et al. 2012]. Já algoritmos como *Gradient Boosting*, *Adaboost* e *XGboost*, como mostrado no estudo [Pessoa et al. 2023], buscam melhorar o desempenho combinando vários modelos fracos em um modelo forte, ajustando repetidamente os pesos dos exemplos de treinamento para corrigir erros anteriores.

A curva ROC, utilizada na avaliação, é um gráfico que representa a taxa de verdadeiros positivos (Sensiblidade) em relação à taxa de falsos positivos (1 - Especificidade) em diferentes pontos de um modelo de classificação. Por sua vez, a especificidade é a proporção de instâncias negativas corretamente identificadas. A AUC (*Area Under The Curve*) é uma medida numérica da qualidade geral do modelo, variando de 0 a 1, onde 1 indica um modelo perfeito e 0.5 indica uma escolha aleatória. Para avaliação, também foi utilizado o *StratifiedKFold*, que é uma técnica valiosa para validação cruzada que proporciona uma avaliação mais equilibrada e confiável dos modelos. Por outro lado, uma tabela que mostra a performance de um modelo de classificação, comparando as previsões do modelo com as classes reais dos dados é chamada matriz de confusão. Ela exibe os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, como considerado no trabalho [Freire et al. 2020].

3. Revisão Bibliográfica

Ainda sobre o Covid-19, o trabalho [Prakash et al. 2020] teve como objetivo identificar o grupo etário mais afetado pela doença, utilizando o *Random Forest Regressor* e o *Random Forest Classifier* como modelos de destaque, superando outros modelos como SVM, KNN (*K-Nearest Neighbors*), *Decision Tree*, *Naive Bayes*, *Multiple Linear Regression*, *Logistic Regression* e *XGboost*. A avaliação dos modelos foi realizada com base na pontuação de regressão *RSquared* e a precisão, dividindo os dados em conjuntos de treinamento(70% dos dados) e teste (30% dos dados). Os resultados revelaram que pessoas entre 20 a 29, 30 a 39 e 40 a 50 anos são os grupos mais afetados pela doença, conforme identificado pelos modelos de previsão. Além disso, os regressores *Random Forest* superaram outros modelos em termos de coeficiente de determinação.

O trabalho [Freire et al. 2020] investigou a demora na detecção do *SARS-CoV-2* devido ao alto custo e tempo do teste PCR. Após dividir os dados em conjuntos de treinamento e teste, selecionaram as 15 características mais importantes com *Random Forest* (RF). Foram treinados modelos SVM (*Support Vector Machine*), MLP (*Multi Layer Perceptron*) e RF, avaliando-os com métricas como precisão, *Recall* e *F1-score*. O modelo RF obteve resultados assertivos para todos os casos de Covid-19 e foi validado com *k-Fold Cross-Validation*. Além disso, a classificação de novas amostras em unidades de internação mostrou-se eficaz na maioria dos casos, destacando-se o modelo MLP na previsão correta de todos os casos negativos para admissão na enfermaria em alguns cenários.

Na pesquisa realizada por [Assaf et al. 2020] foram utilizados *Neural Network*, RF e *Classification and Regression Trees* (CRT) com o objetivo de prever o risco de pacientes desenvolverem um quadro crítico de COVID-19, com base em seu status no momento da admissão. Dessa maneira, constataram que a saturação de oxigênio no ar ambiente foi o melhor preditor isolado, com uma AUC de 0,787. Após dez validações cruzadas, o modelo *Neural Network* obteve a melhor performance. O modelo RF apresentou uma melhora de 12,0% em relação ao APACHE II *score*, enquanto o modelo CRT alcançou sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo 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 a contagem de células sanguíneas, o tempo desde os sintomas até a admissão, a saturação de oxigênio e a contagem de linfócitos no sangue.

4. Experimento

4.1. Análise manual dos dados

Inicialmente, foi feito um estudo observacional de cada coluna da tabela de dados. Procurou-se compreender o significado de cada valor em colaboração com um especialista da área da saúde. Para essa atividade analítica foi empregada a ferramenta *Power BI*, com destaque para a importância no processo de visualização e compreensão dos dados relacionados à área da saúde.

4.2. Preparação e adequação dos dados

No início do processo, os dados foram padronizados e ajustados para lidar com valores incompletos e vazios, a fim de deixar os registros em um mesmo formato. Foram utilizadas técnicas para remover pontuação incorreta e preencher lacunas com zero. A conversão de textos em valores numéricos permitiu uma implementação eficaz de algoritmos de aprendizado de máquina. Para isso, foi utilizada a técnica *bag of words* para extrair informações de colunas com variáveis com subconjuntos semelhantes, uma abordagem estudada mais a fundo no trabalho de [Qader et al. 2019].

Dessa maneira, as listas criadas através da técnica *bag of words* foram convertidas em números binários usando o *MultiLabelBinarizer* (ferramenta Python usada para converter rótulos de várias classes em uma matriz binária), para facilitar o treinamento dos modelos. Foi utilizado o *RandomUnderSampler* (seleciona e deleta aleatoriamente uma amostra da classe majoritária para igualar o número de instâncias da classe minoritária) para equilibrar as classes na coluna de IRAS. Dessa forma foi possível obter um conjunto de dados balanceados onde é visto exatamente as classes iguais: 57 amostras

da classe 0 (não adquiriu IRAS) e mantendo a classe 1 (adquiriu IRAS) no total de 57 amostras também. Como visto no estudo de [Dixit 2022], para evitar viés decorrente da diferença entre as classes, é importante buscar esse equilíbrio nas amostras e criar uma representação mais equilibrada dos resultados.

4.3. Seleção de características

Conforme demonstrado no estudo [Raut et al. 2022], foi utilizada uma técnica de pré-processamento chamada *LabelEncoder*, facilitando o treinamento do modelo de aprendizado de máquina. Esse material é uma classe da biblioteca scikit-learn em Python que é usada para converter rótulos de texto em números. Logo, a coluna sobre IRAS foi selecionada como a variável-alvo, refletindo o foco do estudo em prever essa condição. Após a preparação inicial dos dados, um passo importante foi identificar as características mais relevantes para prever IRAS, utilizando uma técnica chamada *Feature Importance* do *Random Forest*, assim como em [Freire et al. 2020]. A análise revelou que as variáveis ”MICRORGANISMO”, ”TOPOGRAFIA”e ”IDADE”são as mais influentes, como demonstrado na Figura 1.

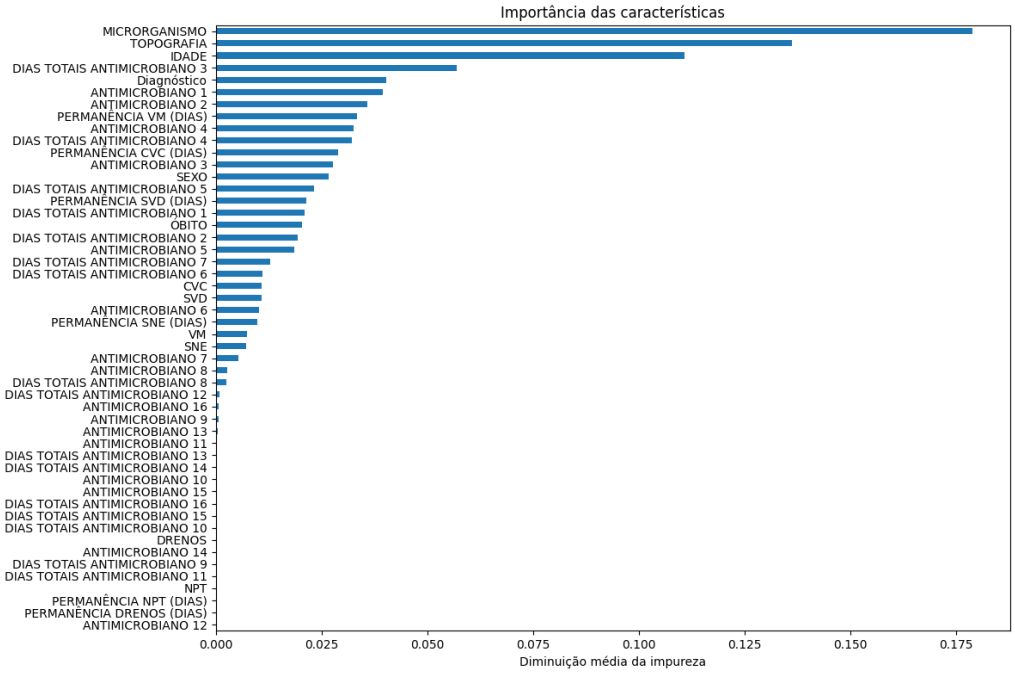

Figura 1. Características mais importantes apresentadas no dados.

4.4. Treinamento e validação dos modelos

Considerando os trabalhos referenciados, a escolha dessa variedade de modelos permitiu uma análise abrangente sobre suas capacidades preditivas em relação às IRAS. Métricas como matriz de confusão, acurácia, precisão, *F1-score*, *Recall* e a curva AUC-ROC foram calculadas, proporcionando uma análise mais ampla da capacidade de previsão das IRAS. Para avaliação dos resultados, foi utilizado 15% do conjunto de dados para teste e 85% para treinamento. A divisão dos dados foi feita dessa maneira devido à quantidade limitada de informações disponíveis para análise. Com um conjunto de dados pequeno, é importante manter uma parte significativa para o treinamento do modelo, garantindo que ele tenha exemplos suficientes para aprender padrões relevantes.

5. Resultados

Após o treinamento e classificação inicial, constatou-se, assim como no trabalho [Prakash et al. 2020], que o modelo *Random Forest* se destacou como o mais eficiente, com resultados de 81% de acurácia, 89% de precisão, *Recall* de 75% e *F1-score* de 80%, como mostrado na Tabela 1. Além disso, foi aplicada a validação cruzada utilizando o método *StratifiedKFold*, que divide o conjunto de treinamento em 5 partes mantendo a proporção das classes, permitindo uma melhor avaliação dos modelos. Na Figura 2 a) é apresentada a matriz de confusão do modelo *Random Forest*, no qual a distribuição foi de 2 falsos negativos, o que indica que o modelo classificou incorretamente dois pacientes como não portadores de IRAS quando, na verdade, eram portadores e 1 falso positivo. Os resultados referentes aos demais modelos estão apresentados na Figura 3.

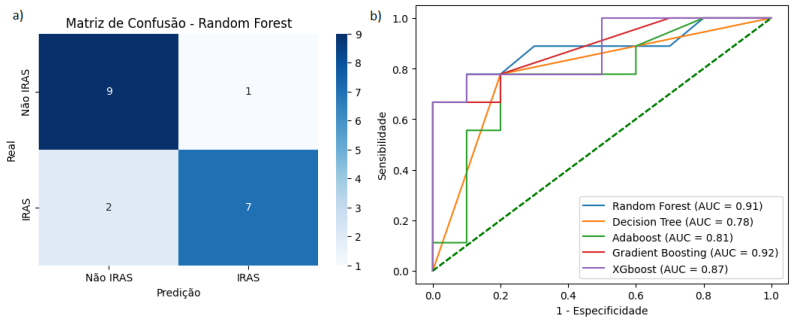

Figura 2. Avaliação dos modelos. a) Matriz de confusão *Random Forest* e b) Curva *Receiver Operating Characteristic* dos modelos.

Tabela 1. Desempenho dos modelos antes do balanceamento de dados.

Modelo	Acurácia	Precisão	*Recall*	*F1-score*
Random Forest	0.8121	0.8912	0.7555	0.8043
Decision Tree	0.7810	0.7727	0.8355	0.7961
Adaboost	0.7594	0.7559	0.7933	0.7705
Gradient Boosting	0.7910	0.8058	0.8133	0.7987
XGboost	0.7905	0.7944	0.8155	0.8001

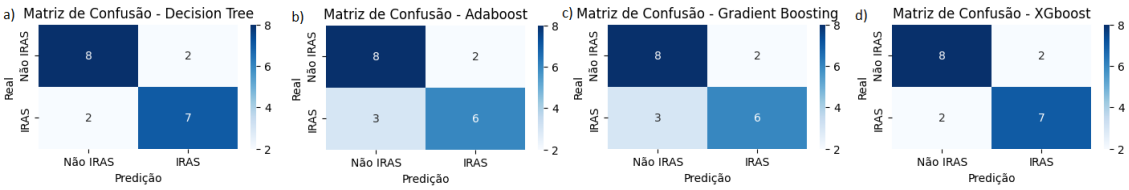

Figura 3. Avaliação dos outros modelos. a) Matriz de confusão *Decision Tree*, b) Matriz de confusão *Adaboost*, c) Matriz de confusão *Gradient Boosting* e d) Matriz de confusão *XGboost*.

Observando a curva ROC exposta na Figura 2 b), é possível verificar como a taxa de verdadeiros positivos se relaciona com a taxa de falsos positivos. Os resultados obtidos mostraram que o *Random Forest* atingiu uma AUC de 0,91 expondo sua capacidade de discriminação entre classes.

6. Considerações Finais

Este estudo explora o uso de algoritmos de aprendizado de máquina para prever Infecções Relacionadas à Assistência à Saúde (IRAS) em pacientes de Unidades de Terapia Intensiva (UTIs). *Random Forest* se destacou por sua eficiência, identificando características mais importantes como "MICRORGANISMO", "TOPOGRAFIA" e "IDADE". As descobertas buscam aprimorar as decisões clínicas e reduzir riscos e complicações associadas a infecções em ambientes hospitalares.

Entretanto, a amostra foi restrita a um único hospital durante um período específico, o que pode limitar a generalização dos resultados. Além disso, a complexidade e a falta de interpretabilidade dos modelos podem ser desafiadoras para profissionais de saúde. O estudo também é exploratório, requerendo pesquisas futuras com bases de dados maiores e abrangendo outros hospitais para validar a eficácia dos modelos em diferentes cenários. Essas limitações ressaltam a necessidade de estudos adicionais para confirmar e expandir as conclusões.

Referências

Ali, J., Khan, R., Ahmad, N., and Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5):272.

Assaf, D., Gutman, Y., Neuman, Y., Segal, G., Amit, S., Gefen-Halevi, S., Shilo, N., Epstein, A., Mor-Cohen, R., Biber, A., et al. (2020). Utilization of machine-learning models to accurately predict the risk for critical covid-19. *Internal and emergency medicine*, 15:1435–1443.

Dixit, R. R. (2022). Predicting fetal health using cardiotocograms: A machine learning approach. *Journal of Advanced Analytics in Healthcare Management*, 6(1):43–57.

Freire, D. L., de Oliveira, R., Carmelo Filho, J., et al. (2020). Machine learning applied in sars-cov-2 covid 19 screening using clinical analysis parameters. *IEEE Latin Am. Trans*, 100(1).

Pessoa, S. M. B., Oliveira, B. S. d. S., Santos, W. G. d., Oliveira, A. N. M., Camargo, M. S., Matos, D. L. A. B. d., Silva, M. M. L., Medeiros, C. C. d. Q., Coelho, C. S. d. S., Andrade Neto, J. d. S., et al. (2023). Predição de choque séptico e hipovolêmico em pacientes de unidade de terapia intensiva com o uso de machine learning. *Revista Brasileira de Terapia Intensiva*, 34:477–483.

Prakash, K. B., Imambi, S. S., Ismail, M., Kumar, T. P., and Pawan, Y. (2020). Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms. *International Journal*, 8(5):2199–2204.

Qader, W., M. Ameen, M., and Ahmed, B. (2019). An overview of bag of words;importance, implementation, applications, and challenges. pages 200–204.

Raut, K., Patil, J., Wade, S., and Tinsu, J. (2022). Mental health and personality determination using machine learning. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, pages 1231–1236. IEEE.