Avaliação do Desempenho de Modelos de Aprendizado de Máquina na Predição da Mortalidade por Sepse em Diferentes Bancos de Dados

Luiza S. B. Leidemer[1], João P. M. Bidart, Marcus V. D. Pavinato, Luciano Z. Goldani, Claúdio F. R. Geyer[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

lsbleidemer@inf.ufrgs.br, geyer@inf.ufrgs.br

Resumo. *A sepse é uma resposta inadequada a infecções, com alta mortalidade, especialmente no Brasil. O estudo visa desenvolver modelos para predizer a mortalidade por sepse, superando métodos tradicionais e avaliando desempenho em bancos de diferentes países. Serão utilizados os algoritmos Decision Tree, Random Forest e XGBoost. Os dados serão coletados de hospitais e bases como MIMIC. Como resultados preliminares, na base MIMIC-IV, o RF apresentou melhor desempenho, com equilíbrio entre precisão, sensibilidade e boa AUC. Na base brasileira, os modelos mostraram overfitting e queda de performance em teste. Espera-se resolver problemas, identificar o melhor modelo, explorar vieses entre bases e avaliar aplicabilidade clínica dos algoritmos.*

Abstract. *Sepsis is an inadequate response to infections, with high mortality, especially in Brazil. This study aims to develop models to predict sepsis mortality, outperforming traditional methods and evaluating performance across databases from different countries. Algorithms such as Decision Tree, Random Forest, and XGBoost will be used. Data will be collected from hospitals and databases such as MIMIC. As preliminary results, in the MIMIC-IV database, RF showed the best performance, balancing precision, recall, and good AUC. In the Brazilian database, the models exhibited overfitting and decreased test performance. The study aims to address problems, identify the best model, explore biases across databases, and assess the clinical applicability of the algorithms.*

1. Introdução

A sepse, ou infecção generalizada, é uma resposta desregulada do organismo a uma infecção, levando à disfunção orgânica progressiva e podendo ser fatal [Instituto Latino-Americano de Sepse 2018]. No Brasil são registrados, anualmente, cerca de 400 mil casos em adultos, com aproximadamente 240 mil óbitos, refletindo uma taxa de mortalidade superior à de países desenvolvidos. A predição da mortalidade por sepse busca identificar precocemente os pacientes em risco, permitindo intervenções preventivas, já que os primeiros momentos da resposta inflamatória são decisivos para o sucesso do tratamento [Brasil. Ministério da Saúde 2023].

Algoritmos de *Machine Learning* (ML) têm sido amplamente empregados para prever a condição, apoiar decisões médicas e identificar pacientes com maior risco de

mortalidade [Bezerra et al. 2023]. Além disso, a aplicação desses modelos permite identificar os pacientes que mais necessitam de suporte médico. Tais modelos podem ser implementados em hospitais com o objetivo de diminuir a mortalidade hospitalar, reduzir o tempo de internação e otimizar os custos de tratamento [Islam et al. 2019].

O estudo visa desenvolver algoritmos de ML para realizar a predição mortalidade por sepse, e avaliar o desempenho dos modelos em bancos de dados de diferentes países, a fim de verificar possíveis vieses dos algoritmos entre bases distintas. Ademais, serão avaliadas as questões éticas da implementação desses modelos no fluxo clínico.

O artigo esta organizado da seguinte forma: Trabalhos relacionados na seção 2, Metodologia na seção 3 e Resultados na seção 4.

2. Trabalhos relacionados

O estudo de [Moor et al. 2021] realizou uma revisão sistemática sobre a predição de sepse em unidades de terapia intensiva (UTIs) utilizando ML. Foram incluídos 22 artigos, a maioria baseada em dados de origem norte-americana. Os autores observaram que diferenças nas características demográficas, nos protocolos clínicos e na disponibilidade de dados afetam a identificação dos casos de sepse, evidenciando um viés de representatividade. Além disso, destacaram a baixa reprodutibilidade dos estudos, não sendo possível identificar o melhor algoritmo de ML, e enfatizaram a necessidade de maior diversidade de dados e validação prospectiva dos modelos.

No artigo de [Chao et al. 2022], os pesquisadores desenvolveram e validaram um modelo de predição de mortalidade hospitalar utilizando diferentes algoritmos de ML. Foram analisadas variáveis demográficas, clínicas, laboratoriais, comorbidades e biomarcadores relacionados à sepse. O RF apresentou o melhor desempenho, superando tanto a regressão logística quanto os escores clínicos tradicionais. Como conclusão, os resultados demonstraram que os modelos de ML fornecem uma predição mais precisa da mortalidade em 28 dias, com maior capacidade de lidar com dados multidimensionais em comparação ao modelo de regressão logística.

3. Metodologia

Este estudo visa prever a mortalidade em 30 dias de pacientes com sepse, usando dados das primeiras 24 horas de internação na UTI. Serão aplicados os algoritmos *Decision Tree* (DT), *Random Forest* (RF) e *Extreme Gradient Boosting* (XGBoost), excluindo redes neurais devido ao tamanho limitado das bases de dados e à necessidade de maior transparência para uso clínico. Duas fontes de dados foram utilizadas: MIMIC-IV (Medical Information Mart for Intensive Care IV) e uma base proveniente de um hospital brasileiro de Porto Alegre. Foram coletados dados demográficos, sinais vitais, exames laboratoriais, comorbidades e intervenções médicas. Incluíram-se pacientes com mais de 18 anos e diagnóstico de sepse. A base brasileira possui 24 variáveis e 404 amostras, enquanto a base MIMIC contém 18 variáveis e 3552 amostras, devido à ausência de dados como escala de coma de Glasgow, índice de comorbidades de Charlson, creatinina, PaO_2, relação PaO_2/FiO_2 e oxigenoterapia. As amostras que possuíam 4 ou mais valores faltantes foram excluídas. .

Os dados foram divididos utilizando validação cruzada estratificada (*Stratified K-Fold*) com cinco *folds*, garantindo que a proporção entre as classes fosse preservada em

cada divisão. Em cada iteração, o conjunto de treino foi separado do teste, e a imputação de valores ausentes foi realizada exclusivamente sobre os dados de treino, utilizando o método *Iterative Imputer* com *Random Forest Regressor* como estimador base. Os valores imputados aprendidos no treino foram aplicados ao conjunto de teste, evitando qualquer vazamento de informação. Considerando o desbalanceamento entre classes, 36% para MIMIC e 17% para a base brasileira, foi aplicada a técnica SMOTE (*Synthetic Minority Over-sampling Technique*).

O modelo *Random Forest* foi configurado com 200 árvores, profundidade máxima de 7, mínimo de 3 amostras por folha, seleção de atributos via raiz quadrada e balanceamento de classes. O *Decision Tree* utilizou profundidade máxima de 5, mínimo de 5 amostras por folha e também aplicou balanceamento de classes. Já o *XGBoost* foi ajustado com 200 estimadores, profundidade máxima de 5, taxa de aprendizado de 0.1, e o parâmetro `scale_pos_weight` igual a 1 para compensar o desbalanceamento. Para cada modelo, foi realizada a otimização do *threshold* de classificação com base no melhor *F1-score*, e as métricas de desempenho foram avaliadas por *fold*, incluindo AUC, F1, *recall*, acurácia e precisão *recall*. Além disso, matrizes de confusão foram elaboradas para permitir uma análise mais detalhada dos erros de classificação.

O estudo seguiu todas as diretrizes éticas para uso de dados secundários, com aprovação do comitê de ética da base brasileira (parecer nº 7.482.646). O uso do MIMIC-IV cumpriu os requisitos éticos do PhysioNet. O Termo de Consentimento Livre e Esclarecido foi dispensado, pois os dados foram obtidos retrospectivamente sem intervenção direta nos pacientes.

4. Resultados

A Tabela 1 apresenta o desempenho médio dos modelos ao longo de cinco *folds*, permitindo a comparação entre os modelos aplicados ao banco de dados MIMIC-IV e à base de dados brasileira.

Tabela 1. Desempenho médio dos modelos nas bases MIMIC-IV e brasileira

	MIMIC IV (Treino)			Base de dados brasileira (Treino)		
	RF	DT	XGBoost	RF	DT	XGBoost
Acurácia	0.836	0.736	0.735	0.987	0.814	1.000
F1-score	0.741	0.594	0.638	0.981	0.561	1.000
Recall	0.642	0.523	0.603	0.982	0.464	1.000
	MIMIC IV (Teste)			Base de dados brasileira (Teste)		
	RF	DT	XGBoost	RF	DT	XGBoost
Acurácia	0.669	0.542	0.663	0.837	0.814	0.829
F1-score	0.607	0.575	0.603	0.562	0.505	0.578
Recall	0.691	0.841	0.689	0.614	0.543	0.671
AUC	0.746	0.705	0.737	0.826	0.695	0.838
Average Precision	0.700	0.613	0.695	0.519	0.384	0.593

Na base MIMIC-IV, os modelos apresentaram desempenho superior no conjunto de treino, com alta acurácia. O RF obteve o maior F1-score (0,741) e *recall* (0,642), indicando melhor equilíbrio entre precisão e sensibilidade. No conjunto de teste, observou-se

redução no desempenho da acurácia e F1-score, com RF mantendo maior acurácia (0,669) e AUC (0,746), enquanto a DT apresentou *recall* elevado (0,841), sugerindo maior sensibilidade à detecção de eventos positivos, porém com menor capacidade de classificação correta dos casos negativos. O aumento no *recall* se dá pelo uso da otimização do *threshold* de classificação com base no melhor *F1-score*.

Na base brasileira, os modelos XGBoost e RF apresentaram desempenho quase perfeito no conjunto de treino, acurácia, F1-score e recall iguais a 1,0, para o XGBoost, e acurácia 0,987, F1-score 0,981, recall 0,982 para RF. Esses resultados indicam forte *overfitting*, sugerindo que os modelos memorizaram as amostras em vez de aprender padrões generalizáveis, comportamento acentuado pelo pequeno tamanho da base (404 amostras). No conjunto de teste, a performance da DT é baixa, mostrando menor capacidade de generalização e maior vulnerabilidade ao pequeno tamanho da amostra. O XGBoost manteve melhor desempenho em teste, com acurácia de 0,829 e *recall* de 0,671, seguido pelo RF (acurácia 0,837, *recall* 0,614), enquanto a DT permaneceu inferior (acurácia 0,814, *recall* 0,543), evidenciando que mesmo modelos robustos sofrem queda de performance em amostras pequenas. Esses resultados reforçam a necessidade de ampliar o tamanho e a diversidade da base para reduzir o *overfitting* e permitir avaliação mais confiável do desempenho dos modelos.

Por fim, mais resultados são esperados, como a identificação do melhor modelo para a predição da mortalidade por sepse; análises das melhores variáveis e suas correlações; e mais bases para análise. A análise de diferentes bancos de dados permitirá identificar potenciais vieses entre as bases e ajustar os modelos conforme necessário. O estudo também avaliará a percepção dos profissionais de saúde sobre o uso do algoritmo, explorando sua integração como ferramenta de apoio à prática clínica.

Referências

Bezerra, A., Maciel, N., Filho, L., Mendes, A., Gois, F., and Silva, L. (2023). Efetividade de algoritmos de inteligência artificial para predição de sepse em adultos de unidades de terapia intensiva: revisão de escopo. *Revista Interfaces*.

Brasil. Ministério da Saúde (2023). Dia mundial da sepse: Ministério da saúde alerta para a importância do diagnóstico precoce. Disponível em: `https://www.gov.br/saude/pt-br/assuntos/noticias/2022`. Acesso: 13 nov. 2024.

Chao, H.-Y., Wu, C.-C., Singh, A., Shedd, A., Wolfshohl, J., Chou, E. H., Huang, Y.-C., and Chen, K.-F. (2022). Using machine learning to develop and validate an in-hospital mortality prediction model for patients with suspected sepsis. *Biomedicines*, 10(4):802.

Instituto Latino-Americano de Sepse (2018). Implementação de protocolo gerenciado de sepse. Disponível em: `https://ilas.org.br/wp-content/uploads/2022/02/protocolo-de-tratamento.pdf`. Acesso 13 nov. 2024.

Islam, M., Nasrin, T., Walther, B. A., Wu, C.-C., Yang, H.-C., and Li, Y.-C. (2019). Prediction of sepsis patients using machine learning approach: A meta-analysis. *Commputer Methods and Programs in Biomedicine*.

Moor, M., Rieck, B., Horn, M., Jutzeler, C. R., and Borgwardt, K. (2021). Early prediction of sepsis in the icu using machine learning: A systematic review. *Frontiers in Medicine*, Volume 8 - 2021.