UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXIX SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2025

Avaliação de métodos de aprendizado de máquina na predição da deterioração clínica de adultos em hospitais

Vanderleicio Carvalho Leite Junior[1]; Matheus Giovanni Pires[2]
1.Vanderleicio Carvalho Leite Junior, PROBIC/UEFS, ENGENHARIA DE COMPUTAÇÃO, e-mail: vanderleiciojr397@gmail.com
2.Matheus Giovanni Pires, DEPARTAMENTO DE CIÊNCIAS EXATAS, e-mail: mgpires@uefs.br

PALAVRAS-CHAVE: Predição, deterioração clínica, aprendizado de máquina,.

INTRODUÇÃO

A deterioração clínica é vista como a condição na qual pacientes internados em unidades hospitalares têm uma piora no seu estado clínico, que leva a uma descompensação fisiológica, podendo prolongar sua estadia no ambiente hospitalar, necessitando de um cuidado mais intensivo, ou mesmo levá-los a morte (Padilla et al., 2018). Estudos já mostram que é possível antecipar esse quadro de piora clínica através do monitoramento de anormalidades nos sinais vitais dos pacientes, minutos ou até horas antes do acontecimento do quadro de piora, e a intervenção pode ajudar a salvar a vida do paciente (Correia & Mayo 2014). Com a criação e o aprimoramento dos sistemas de armazenamento de dados, cada vez mais informações clínicas têm sido coletadas e armazenadas, o que permite que modelos de inteligência artificial sejam usados para trabalhar nesses dados em diferentes estratégias (Solís-García et al., 2023).

Diante dessa necessidade do uso de dados clínicos para pesquisas e, principalmente, para pesquisas voltadas ao uso de modelos de *machine learning* (ML), tem-se a necessidade da determinação de um *dataset* que reúna as informações necessárias para a tarefa apresentada. Dentre os principais conjuntos de dados dessa área que estão disponíveis publicamente, destaca-se o *Medical Information Mart for Intensive Care* (MIMIC), que reúne dados de pacientes internados em Unidades de Tratamento Intensivo (UTI) do Beth Israel Deaconess Medical Center, e permite acessar informações desidentificadas sobre sinais vitais, exames laboratoriais, remédios administrados etc., bem como dados demográficos dos pacientes (sem permitir o rastreio deles) (Johnson et al., 2023).

Considerando o contexto apresentado e a disponibilidade das informações clínicas pelo *dataset* MIMIC-IV, este trabalho tem o objetivo de avaliar o uso de modelos de *machine learning* para prever, com base em dados clínicos e demográficos dos pacientes internados em UTI, a possibilidade de eles evoluírem para um quadro de piora clínica.

METODOLOGIA

Para construir os conjuntos de dados que foram usados para treinar e validar os modelos partiu-se das informações presentes no já citado MIMIC-IV na sua versão v3.1. O MIMIC separa as informações em dois blocos principais o “hosp” - que contém informações sobre

toda a estadia hospitalar do paciente, com foco em dados demográficos e laboratoriais - e no "icu" - que contém dados mais específicos das estadias dos pacientes nas UTIs do hospital, com foco nos eventos que aconteceram durante a estadia. As tabelas do MIMIC das quais os dados usados foram retirados são: "icustays" e "admissions", com as informações acerca da chegada e saída do paciente da UTI e informações referente ao óbito; "patients", com as informações de gênero e idade dos pacientes; e a principal "chartevents", com todos os eventos que foram registrados durante a permanência do paciente na UTI, contendo tanto sinais vitais quanto resultados de alguns exames laboratoriais dos pacientes. A Tabela 1 mostra a lista dos sinais vitais e exames que foram usados para compor os dados dos experimentos.

Tabela 1. Sinais vitais e resultados de exames usados.

Nome no MIMIC-IV	Tipo	Nome no MIMIC-IV	Tipo
Gender	Categorical	Heart Rate	Numeric
Age	Numeric	Level of Consciousness	Categorical
Daily Weight	Numeric	PAR-Respiration	Categorical
Respiratory Rate	Numeric	Glucose (whole blood)	Numeric
O2 saturation pulseoxymetry	Numeric	Sodium (whole blood)	Numeric
Temperature Celsius	Numeric	Creatinine (whole blood)	Numeric
Arterial Blood Pressure systolic	Numeric	Potassium (whole blood)	Numeric
Arterial Blood Pressure diastolic	Numeric		

Após a escolha dos dados fisiológicos que seriam capturados, partiu-se para a extração deles e formação dos *datasets*. Primeiramente foram selecionados pacientes que tivessem pelo menos 4 dias de estadia na UTI, o que permite avaliar os modelos de ML dentro de um cenário de permanência maior de pacientes em UTI. Foram consideradas duas maneiras de organizar os dados, a primeira, proposta por (Cheng et al., 2020), forma um vetor diário do paciente, constando os três últimos valores coletados de cada sinal ou exame do paciente; já a outra configuração, usada em (Baker et al., 2020), propõe determinar 7 medidas estatísticas para cada sinal ou exame, também diariamente, as medidas são: primeiro valor, último, maior, menor, média dos valores, mediana e o desvio padrão; em ambas as formas a classificação de cada vetor formado é dada pela piora ou não do paciente, no caso deste trabalho se ele foi ou não a óbito no fim do dia representado.

Embora o MIMIC-IV represente um resultado robusto para seu propósito, ele ainda pode apresentar algumas falhas, que são inerentes ao seu processo de formação. Não necessariamente todos os sinais vitais e exames laboratoriais são solicitados e/ou coletados a todo momento, por isso, os vetores diários formados apresentaram valores faltantes, ou seja, posições em que os dados não tinham sido registrados ou até mesmo nem existiam. Com o intuito de mitigar o efeito desse problema nos modelos de ML usados, cinco algoritmos de imputação de dados avaliados em (Solís-García et al., 2023) foram implementados, são eles: *linear interpolation*, *indicator imputation*, *forward filling*, *carry forward* e *zero imputation*.

Para a realização da tarefa foram considerados 3 modelos de *machine learning* o LightGBM, o eXtreme Gradient Boosting (XGBoost) e o Random Forest Classifier,

sendo os dois primeiros implementados em sua versão de classificador das bibliotecas próprias e o último a partir da biblioteca Scikit-learn, todos em Python. Os modelos foram tiveram seus parâmetros afinados utilizando o algoritmo *"grid search"* onde um conjunto de configurações são treinadas e testadas e a que teve o melhor desempenho é escolhida. Assim, foram formados no total 18 experimentos, cada um representado pela junção de um tipo de modelo de ML, com um dos conjuntos de dados, variando a forma de imputação, totalizando 15 configurações e mais 3 configurações vindas da junção dos modelos com a organização baseada em estatísticas dos dados daquele dia.

Devido a natureza desbalanceada do problema - foram 220286 de vetores de dias em que não houve óbitos, comparados a 3842 de dias em que houve - métricas como acurácia não refletem bem o desempenho dos modelos, por isso, para escolher qual configuração seria escolhida, bem como comparar os modelos, foi utilizada a métrica *G-mean*, calculada por $Gmean = \sqrt{sensibilidade \,.\, especificidade}$, o que balanceia os acertos dos exemplos da classe positiva e da negativa.

RESULTADOS E DISCUSSÃO

Diante deste cenário, todas as 18 configurações propostas foram avaliadas através da técnica de validação cruzada com 10 *folds*, onde para cada *fold* os dados do treino foram balanceados utilizando o algoritmo *Random Under Sample*, assim os modelos aprenderiam entendendo que ambas as classes têm o mesmo nível de importância, mas são testados utilizando os dados desbalanceados, refletindo a situação real.

A Tabela 3 traz os resultados das principais métricas calculadas para as configurações que tiveram os cinco melhores *G-means* dentre as 18 avaliadas. As configurações foram nomeadas de acordo com os dados usados ("sta": medidas estatísticas dos dados diários; "zr": uso do *zero imputation*; "id": uso do *indicator imputation*), seguido do modelo que foi usado ("RF": Random Forest Classifier; "LGBM": LightGBM; "XGB": XGBoost), o maior valor de cada coluna foi destacado em negrito e o menor com um sublinhado.

Tabela 2. Média e desvio padrão das métricas calculadas para os 10 *folds*.

Configuração	G-mean (méd)	G-mean (dp)	Sens. (méd)	Sens. (dp)	Espec. (méd)	Espec. (dp)
staRF	**0.9553**	0.0011	**1.000**	0.0000	**0.9127**	0.0021
zrRF	0.9497	0.0012	**1.000**	0.0000	0.9019	0.0023
idRF	0.9461	0.0020	0.9984	0.0017	0.8965	**0.0033**
staLGBM	0.9409	**0.0038**	0.9792	**0.0075**	0.9041	0.0020
staXGB	0.9330	**0.0038**	0.9643	0.0073	0.9027	0.0018

Legenda: Sens: Sensibilidade; Espec: Especificidade; méd: média; dp: desvio padrão.

Esses resultados indicam que modelos baseados no algoritmo Random Forest parecem se destacar ao tentar prever a mortalidade dos pacientes internados em UTIs, mesmo que a forma que os dados são imputados e organizados varie. Já os modelos LightGBM e XGBoost se beneficiaram mais de organizar os dados considerando variações estatísticas dos sinais vitais dos pacientes em cada um dos dias. Para validar as diferenças dessas configurações, foi feita uma análise estatística usando o teste de Friedman com Nemenyi post hoc e um nível de significância de 0.05 entre os resultados

nos 10 *folds* das configurações. As configurações apresentadas não apresentaram diferença estatística significativa, indicando que embora cenários baseados no uso de Random Forest tenham aparecido em ranques melhores em termos de valor médio do G-mean, estatisticamente eles são semelhantes ao uso da organização estatística com os outros modelos.

CONSIDERAÇÕES FINAIS

A aplicação de modelos de *machine learning* em dados clínicos têm se mostrado promissora, e a tarefa de tentar prever com antecedência a piora clínica de pacientes internados em UTIs pode ser uma das aplicações para agregar o uso de inteligências artificiais na área da saúde. Este trabalho se propôs a avaliar como os modelos Random Forest, LightGBM e XGBoost desempenham ao tentar realizar a tarefa supracitada, unindo os modelos com formas diferentes de organizar os dados e de resolver os problemas de balanceamento e ausência de dados. Os resultados se mostraram promissores, os modelos conseguiram ter um desempenho consideravelmente bom quando avaliados com o G-mean, embora não tenha havido diferença estatística significativa entre as cinco melhores configurações levantadas.

REFERÊNCIAS

BAKER, Stephanie; XIANG, Wei; ATKINSON, Ian. Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach. **Scientific Reports**, v. 10, n. 1, 2020.

CHENG, Fu-Yuan; JOSHI, Himanshu; TANDON, Pranai; *et al*. Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients. **Journal of Clinical Medicine**, v. 9, n. 6, p. 1668, 2020.

CORREIA, Nuno; RODRIGUES, Rui Paulo; SÁ, Márcia Carvalho; *et al*. Improving recognition of patients at risk in a Portuguese general hospital: results from a preliminary study on the early warning score. **International Journal of Emergency Medicine**, v. 7, n. 1, 2014.

JOHNSON, Alistair E. W.; BULGARELLI, Lucas; SHEN, Lu; *et al*. MIMIC-IV, a freely accessible electronic health record dataset. **Scientific Data**, v. 10, n. 1, 2023.

PADILLA, Ricardo M; MAYO, Ann M. Clinical deterioration: A concept analysis. **Journal of Clinical Nursing**, v. 27, n. 7–8, p. 1360–1368, 2018.

SOLÍS-GARCÍA, Javier; VEGA-MÁRQUEZ, Belén; NEPOMUCENO, Juan A.; *et al*. Comparing artificial intelligence strategies for early sepsis detection in the ICU: an experimental study. **Applied Intelligence**, v. 53, n. 24, p. 30691–30705, 2023.