Análise Estatística e Comparativa de Algoritmos de Machine Learning para Predição de Mortalidade em Unidades de Terapia Intensiva

Statistical and Comparative Analysis of Machine Learning Algorithms for Mortality Prediction in Intensive Care Units

Regisson César Pereira de Aguiar[1]
orcid.org/0009-0003-4482-2639

Roberta Andrade de Araújo Fagundes[2]
orcid.org/0000-0002-7172-4183

[1]Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: rcpa@poli.br

[2]Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: roberta.fagundes@upe.br

DOI: 10.25286/repa.v11i3.3974

RESUMO

A predição de mortalidade em unidades de terapia intensiva (UTI) é um desafio crítico para a medicina. Este trabalho propõe a aplicação de algoritmos de Machine Learning para estimar o risco de óbito hospitalar, utilizando dados da base MIMIC-III processados com uma estratégia de interpolação inspirada no sistema MedLens. Foram avaliados cinco modelos supervisionados, com destaque para técnicas de ensemble como LightGBM e XGBoost. Diferentemente de abordagens anteriores, implementou-se um protocolo de validação rigoroso com 25 execuções independentes, seguido de testes de hipótese. Os resultados comprovaram a superioridade estatística do LightGBM, que atingiu acurácia média de 92,99% e F1-Score de 0,927. A aplicação do teste de Wilcoxon Signed-Rank confirmou ($p < 0,05$) que o desempenho do LightGBM é significativamente superior aos demais competidores, validando sua robustez como ferramenta de suporte à decisão clínica e otimização de recursos hospitalares.

PALAVRAS-CHAVE: *Machine Learning, Predição de Mortalidade, Validação Estatística, LightGBM, MIMIC-III.*

ABSTACT

Mortality prediction in intensive care units (ICUs) is a critical challenge for medicine. This work proposes the application of Machine Learning algorithms to estimate the risk of in-hospital death, using data from the MIMIC-III database processed with an interpolation strategy inspired by the MedLens system. Five supervised models were evaluated, with an emphasis on ensemble techniques such as LightGBM and XGBoost. Unlike previous approaches, a rigorous validation protocol was implemented with 25 independent runs, followed by hypothesis testing. The results proved the statistical superiority of LightGBM, which achieved an average accuracy of 92.99% and an F1-Score of 0.927. The application of the Wilcoxon Signed-Rank test confirmed ($p < 0.05$) that LightGBM's performance is significantly superior to other competitors, validating its robustness as a tool for clinical decision support and hospital resource optimization.

KEY-WORDS: *Machine Learning, Mortality Prediction, Statistical Validation, LightGBM, MIMIC-III.*

REVISTA DE Engenharia e Pesquisa Aplicada

1 INTRODUÇÃO

Sinais vitais médicos são medidas essenciais que indicam a temperatura corporal, frequência cardíaca e pressão arterial de um indivíduo, refletindo seu estado geral de saúde. Além desses sinais médicos, os registros mais comuns consistem em anotações clínicas em formato textual, redigidas por médicos ou enfermeiros. O estado de saúde de um paciente apresenta forte correlação com os padrões variáveis presentes nesses sinais médicos. Diversos estudos concentram-se exclusivamente no uso desses sinais para a predição de mortalidade **[1]**.

É crucial que os médicos sejam capazes de prever com precisão a mortalidade de pacientes em Unidades de Terapia Intensiva (UTI), pois isso permite a elaboração de planos de tratamento mais eficazes e avaliações de risco mais acuradas, aumentando as chances de sobrevivência dos pacientes. A análise e interpretação manual de grandes volumes de dados clínicos são tarefas onerosas e dependem fortemente da expertise de especialistas. Por essa razão, técnicas de Machine Learning (Aprendizado de Máquina) têm sido introduzidas para realizar análises automáticas e apoiar a tomada de decisão inteligente **[2]**.

No entanto, a maioria das pesquisas anteriores focou no uso de algoritmos avançados de Machine Learning ou Deep Learning para modelar anotações clínicas com o objetivo de alcançar predições precisas, sem mensurar ou caracterizar adequadamente as anotações originais. Problemas como a frequência irregular de registros, altas taxas de dados ausentes (missing data) e a grande variedade de tipos de sinais limitam significativamente o desempenho da predição de mortalidade, mesmo quando utilizadas técnicas avançadas de inteligência artificial **[2]**.

Diante deste cenário, o presente trabalho propõe a aplicação de modelos inteligentes para prever taxas de mortalidade utilizando técnicas de Machine Learning aplicadas a sinais vitais e anotações clínicas, com o objetivo de desenvolver modelos preditivos precisos. Esta abordagem busca não apenas aprimorar a precisão das predições, mas também discutir as implicações dos resultados para a saúde pública e para políticas preventivas.

2 OBJETIVOS

O objetivo geral deste trabalho é avaliar e comparar estatisticamente a eficácia de diferentes paradigmas de Machine Learning na predição de mortalidade hospitalar de pacientes em Unidades de Terapia Intensiva, utilizando dados da base MIMIC-III **[3]** processados pela metodologia MedLens.

Para alcançar este propósito, definiram-se os seguintes objetivos específicos:

- Implementar e avaliar o desempenho preditivo de cinco algoritmos de classificação supervisionada: Extreme Gradient Boosting (XGBoost) **[4]**, Light Gradient Boosting Machine (LightGBM) **[5]**, Adaptive Boosting (AdaBoost) **[6]**, Support Vector Classifier (SVC) Calibrado **[7,8]** e Kernel Ridge Regression **[9]**. Selecionados por representarem o estado da arte em dados tabulares.
- Aplicar técnicas de pré-processamento avançado, incluindo seleção de características e interpolação de séries temporais, para mitigar os problemas de esparsidade e irregularidade típicos de dados clínicos reais.
- Estabelecer um protocolo de validação robusto, executando 25 rodadas independentes de treinamento e teste para cada modelo, visando mensurar a estabilidade e a variância dos resultados.
- Realizar testes de hipótese formal (Shapiro-Wilk e Teste t pareado) para determinar se as diferenças de acurácia observadas entre os modelos líderes (LightGBM e XGBoost) são estatisticamente significativas, superando as limitações de comparações baseadas apenas em médias simples.

3 CLASSIFICADORES UTILIZADOS

Este estudo avalia cinco algoritmos de aprendizado supervisionado, selecionados por sua eficácia comprovada em dados tabulares e pela diversidade de paradigmas de aprendizado que representam:

- **XGBoost**: O Extreme Gradient Boosting é uma implementação otimizada do Gradient Boosting focada em velocidade, regularização e escalabilidade. Proposto por Chen e Guestrin **[4]**, o algoritmo se destaca pelo tratamento eficiente de dados esparsos e pelo uso de processamento paralelo na construção das árvores, sendo amplamente considerado o estado da arte para dados estruturados.
- **LightGBM**: Desenvolvido pela Microsoft, o Light Gradient Boosting Machine introduziu técnicas como Gradient-based One-Side

Sampling (GOSS) e Exclusive Feature Bundling (EFB). Segundo Ke et al. **[5]**, essas inovações permitem reduzir drasticamente o tempo de treinamento e o uso de memória sem comprometer a acurácia, tornando-o ideal para grandes volumes de dados.

- **AdaBoost**: O Adaptive Boosting, introduzido por Freund e Schapire **[6]**, é um meta-algoritmo seminal que constrói um classificador forte a partir de uma combinação linear de classificadores fracos. A cada iteração, o modelo ajusta os pesos das instâncias, penalizando aquelas classificadas incorretamente nas etapas anteriores para forçar o aprendizado nos casos mais complexos.
- **SVC Calibrado**: O Support Vector Classifier (SVC) baseia-se na teoria de Statistical Learning estabelecida por Cortes e Vapnik **[7]**, buscando encontrar o hiperplano que maximiza a margem de separação entre as classes. Como o SVC padrão não produz probabilidades nativas, utiliza-se o método de Platt Scaling, proposto por Platt [8], para calibrar as saídas do modelo e fornecer estimativas probabilísticas de risco confiáveis.
- **Kernel Ridge**: Este método combina a regularização da Regressão Ridge (norma L2) com o "truque de kernel" (kernel trick). Conforme descrito por Saunders et al. **[9]**, essa abordagem permite que o algoritmo aprenda funções lineares em espaços de características de alta dimensão, capturando relações não-lineares complexas com um custo computacional inferior ao de redes neurais para datasets de tamanho moderado.

3.1 Trabalhos Relacionados

Foram analisados trabalhos publicados no período de 2018 a 2024 que investigam a aplicação de algoritmos de Machine Learning para a predição de desfechos clínicos em unidades de terapia intensiva, com ênfase particular na utilização da base de dados MIMIC-III e na comparação com métodos estatísticos tradicionais.

Estudos recentes têm demonstrado o potencial de modelos baseados em registros eletrônicos de saúde (EHR) para a estratificação de risco em pacientes críticos crônicos. Pesquisadores em **[10]** focaram especificamente em pacientes com diabetes tipo 2 admitidos em UTIs, utilizando um subconjunto de 14.145 registros do MIMIC-III. O objetivo foi prever a mortalidade em múltiplos horizontes temporais (3, 30 e 365 dias) e a readmissão hospitalar. Dentre os algoritmos testados (incluindo Naive Bayes, Regressão Logística e SVC), o AdaBoost e o Multilayer Perceptron (MLP) apresentaram os melhores desempenhos, alcançando AUCs de até 0,81 e acurácia de 88,3%, evidenciando a eficácia de técnicas de aprendizado supervisionado no suporte à decisão clínica.

A superioridade do Machine Learning sobre sistemas de pontuação tradicionais também foi verificada em **[11]**, que utilizou dados de 12.377 pacientes de UTI em Taiwan. O estudo comparou modelos como Random Forest e XGBoost contra escores clínicos clássicos, como SOFA, APACHE II e SAPS II. Os modelos computacionais demonstraram desempenho significativamente superior, com o XGBoost atingindo uma AUC de 0,88. O estudo destacou a fração inspirada de oxigênio (FiO2) na admissão como um fator preditivo crítico, reforçando o papel do ML na otimização de decisões terapêuticas precoces.

Na mesma linha, autores em **[12]**, desenvolveram um modelo preditivo utilizando o algoritmo XGBoost em uma coorte de 5.211 pacientes de UTIs clínicas e cirúrgicas. O modelo superou sistemas tradicionais e a regressão logística, obtendo valores de AUC entre 0,93 e 0,98. Um achado relevante deste estudo foi a importância das anotações de enfermagem (como a escala de agitação e sedação RASS) como fontes confiáveis de dados para a predição de risco, sugerindo que dados não estruturados ou subjetivos possuem alto valor preditivo quando bem processados.

Por fim, abordando o desafio da qualidade dos dados em saúde, o trabalho proposto em **[2]** desenvolveu o sistema MedLens utilizando a base MIMIC-III. Os autores identificaram que muitos sinais vitais possuem altas taxas de valores ausentes e baixa correlação linear com a mortalidade, o que compromete modelos preditivos convencionais. Para mitigar esse problema, o MedLens implementa uma seleção automática de sinais relevantes baseada em medidas estatísticas e uma abordagem de interpolação flexível para séries temporais. Após o tratamento dos dados, o uso de classificadores ensemble resultou em uma performance elevada, com AUC-ROC de 0,96, superando benchmarks anteriores. Esta metodologia de pré-processamento e seleção de características serve de base para a abordagem adotada no presente artigo.

DOI: 10.25286/repa.v11i3.3974

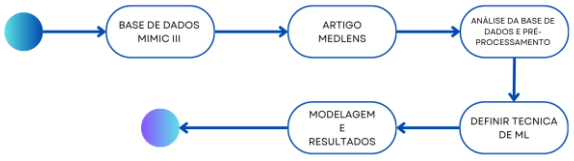
4 METODOLOGIA

Este estudo foi inspirado na abordagem proposta pelo artigo "MedLens: Improve Mortality Prediction Via Medical Signs Selecting and Regression" **[2]**, que aborda a predição de mortalidade de pacientes através da seleção inteligente de sinais vitais e da interpolação de séries temporais. O MedLens foi projetado para selecionar automaticamente os sinais médicos vitais mais discriminativos e utilizar uma abordagem de interpolação flexível para lidar com séries temporais que apresentam altas taxas de dados ausentes (missing data).

A estratégia de pré-processamento foi inspirada na arquitetura MedLens **[2]**, especificamente no módulo de tratamento de dados ausentes. Diferentemente da implementação original, que propõe um pipeline completo de ponta a ponta, este estudo foca na adaptação da técnica de Interpolação por Regressão (Regression-based Interpolation). Para tal, utilizou-se um Random Forest Regressor para estimar e imputar valores faltantes nas séries temporais dos sinais vitais, explorando a correlação não-linear entre as variáveis observadas para reconstruir a dinâmica temporal dos pacientes antes da classificação.

Nosso objetivo é aplicar uma metodologia semelhante para aprimorar a predição de mortalidade utilizando a base de dados MIMIC-III. Problemas de qualidade dos dados, como a alta taxa de ausência em diversos sinais vitais, são tratados através da seleção cuidadosa dos sinais mais relevantes e da aplicação de técnicas de interpolação para preenchimento de valores faltantes. O fluxo metodológico completo adotado neste trabalho pode ser visualizado na Figura 1.

Figura 1 – Metodologia do Artigo



Fonte: Autores.

Diferentemente da abordagem original, que focava em um conjunto restrito de modelos, este trabalho expande a avaliação para cinco algoritmos de aprendizado supervisionado de alto desempenho: Extreme Gradient Boosting (XGBoost) **[4]**, Light Gradient Boosting Machine (LightGBM) **[5]**, Adaptive Boosting (AdaBoost) **[6]**, Support Vector Classifier (SVC) Calibrado **[7,8]** e Kernel Ridge Regression **[9]**. O uso de classificadores de ensemble (como os baseados em Gradient Boosting) e métodos de kernel visa aumentar a acurácia e garantir robustez computacional, buscando atingir ou superar a performance reportada pelos trabalhos anteriores.

Para garantir que as comparações entre os classificadores não sejam enviesadas por flutuações aleatórias na amostragem dos dados, estabeleceu-se um protocolo rigoroso de validação experimental:

- **Múltiplas Execuções**: Cada um dos cinco algoritmos foi submetido a 25 execuções independentes. Em cada execução, os dados foram reamostrados aleatoriamente em conjuntos de treinamento e teste, garantindo que as métricas de desempenho (Acurácia, F1-Score e AUC) reflitam a estabilidade do modelo e não apenas um "caso de sorte" de uma única divisão de dados.
- **Verificação de Normalidade**: As distribuições de acurácia resultantes das 25 rodadas foram submetidas ao teste de Shapiro-Wilk para verificar a aderência à normalidade.
- **Teste de Hipótese**: Para comparar os modelos de melhor desempenho (identificados como LightGBM e XGBoost), aplicou-se o Teste t de Student Pareado (Paired t-test) com nível de significância de 5% ($\alpha = 0.05$). Este teste determina se a diferença média de desempenho entre os modelos é estatisticamente significativa, fornecendo uma base sólida para a seleção do melhor classificador.

4.1 MIMIC-III Database

Este estudo utiliza a base de dados MIMIC-III (Medical Information Mart for Intensive Care III), amplamente empregada em pesquisas de terapia intensiva. O conjunto de dados contém informações clínicas anonimizadas de mais de 40.000 pacientes admitidos em UTIs entre 2001 e 2012, abrangendo sinais vitais, exames laboratoriais, tempo de permanência, diagnósticos e desfechos hospitalares **[3]**.

Foram selecionadas as tabelas CHARTEVENTS e LABEVENTS, que contêm registros de sinais vitais e resultados laboratoriais, respectivamente. Para viabilizar a análise computacional, a base de dados foi reduzida por meio de uma amostragem estratificada **[13]**, mantendo 30% dos registros

originais e preservando a representatividade por tipo de variável.

Aplicou-se uma filtragem baseada em entropia para identificar e remover registros com baixo teor informativo. A entropia, que mensura a incerteza ou variabilidade dos dados, foi utilizada como critério de corte: definiu-se um limiar para remover aproximadamente 30% dos registros com menor variação informativa, retendo aqueles com maior variabilidade para preservar a integridade informacional do conjunto de dados.

A amostragem estratificada é preferível em estudos estatísticos, pois melhora a representatividade ao dividir a população em subgrupos homogêneos, reduz a variabilidade dos resultados e aumenta a eficiência estatística ao otimizar a coleta em cada estrato. Essa abordagem assegura resultados mais equitativos e facilita análises subsequentes, permitindo uma interpretação mais precisa das relações entre as variáveis **[13]**.

Para a constituição do conjunto de dados experimental, adotou-se um protocolo de redução de dimensionalidade baseado em Amostragem Estratificada (*Stratified Sampling*), selecionando-se 30% do volume total de registros das tabelas CHARTEVENTS e LABEVENTS. Esta abordagem, fundamentada nos princípios de delineamento experimental de Montgomery **[13]**, garante que a distribuição original das classes (desfechos de mortalidade e sobrevivência) e a variância das variáveis preditoras sejam preservadas na amostra, assegurando sua representatividade estatística em relação à população total do MIMIC-III.

A decisão por trabalhar com este subconjunto não se deve apenas à eficiência computacional, mas também à maximização da qualidade da informação. Aplicou-se um filtro de entropia para remover registros com baixa variabilidade informativa (ruído ou redundância), conforme metodologia sugerida por Ye e Wu **[2]** no framework MedLens. Estudos prévios demonstram que modelos treinados em subconjuntos de dados limpos e balanceados frequentemente superam ou igualam o desempenho de modelos treinados em dados brutos massivos, devido à redução do overfitting causado por observações ruidosas.

Além disso, a validade desta amostragem é corroborada pelo protocolo de validação deste estudo: a estabilidade dos resultados (baixo desvio padrão) observada ao longo de 25 execuções independentes com reamostragem aleatória confirma que o subconjunto de 30% é suficiente para capturar os padrões preditivos subjacentes sem introduzir viés de seleção.

4.2 Análise Exploratória dos Dados

Para fundamentar a modelagem preditiva, realizou-se uma análise exploratória detalhada sobre a tabela CHARTEVENTS do MIMIC-III, que concentra os registros de sinais vitais. Devido à volumetria massiva dos dados brutos, o processamento foi realizado em lotes (chunks) utilizando a biblioteca Pandas com paralelização via ThreadPoolExecutor.

4.2.1 Estatística Descritiva e Definição de Variáveis

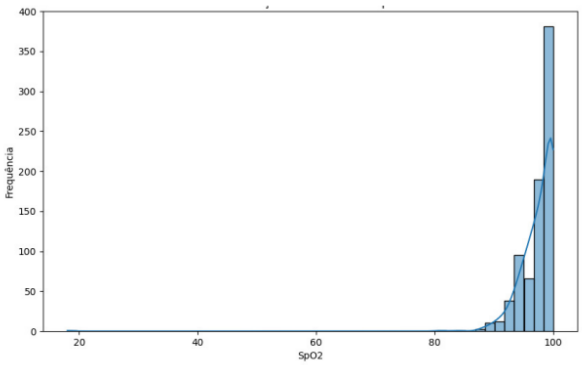
A inspeção das colunas principais revelou a estrutura e a variabilidade dos dados. Atendendo à necessidade de clareza clínica, as variáveis analisadas são definidas a seguir:

- **ROW_ID**: Identificador único de cada linha de registro. A alta cardinalidade (variando de 1 a 330.712.483 na base completa) reflete a granularidade dos eventos monitorados.
- **SUBJECT_ID**: Identificador único do paciente. A concentração de valores indica que múltiplos registros estão associados a um mesmo indivíduo, permitindo a construção do histórico clínico.
- **HADM_ID**: Identificador da admissão hospitalar. Essencial para agrupar eventos ocorridos durante um episódio específico de internação.
- **ITEMID**: Código que define o tipo de medição (ex: Frequência Cardíaca, pH, SpO2).
- **CHARTTIME**: Timestamp do registro, fundamental para a ordenação das séries temporais.

A análise de valores ausentes (missing values) indicou que campos estruturais como ICUSTAY_ID (identificador da estadia na UTI) e VALUEUOM (unidade de medida) são consistentemente preenchidos. Em contrapartida, colunas como RESULTSTATUS e STOPPED apresentaram alta taxa de ausência, o que é esperado dada a natureza de sinalização esporádica dessas variáveis, não impactando a análise clínica principal.

Figura 2 – Distribuição dos Valores de Saturação Periférica de Oxigênio (SpO2)

DOI: 10.25286/repa.v11i3.3974



Fonte: Autores.

A análise de dispersão (Figura 2) das variáveis numéricas (vide Fig. 2 do material suplementar) revelou comportamentos distintos. A variável VALUENUM, que armazena os valores clínicos aferidos, apresenta outliers significativos, mas com a massa de dados concentrada em intervalos fisiológicos esperados. As variáveis de alerta (WARNING e ERROR) apresentaram variância quase nula, indicando que a base, apesar de complexa, possui alta confiabilidade nos registros validados.

A distribuição dos valores de Saturação Periférica de Oxigênio (SpO2) demonstra consistência com a realidade clínica de terapia intensiva. A maioria das medições concentra-se no intervalo de 90% a 100%, com pico em 100%, refletindo pacientes sob monitorização e suporte de oxigênio.

- **95% - 100%**: Faixa esperada para indivíduos saudáveis ou estabilizados.
- **91% - 94%**: Indicativo de hipoxemia leve, exigindo monitoramento para prevenir agravamento.
- **86% - 90%**: Hipoxemia moderada (dispneia, taquicardia), demandando intervenção médica.
- **< 85%**: Hipoxemia grave (risco de cianose e danos orgânicos), necessitando de socorro imediato **[14]**.

A escassez de registros abaixo de 80% é coerente, representando episódios críticos raros ou eventuais erros de leitura do sensor. Essa distribuição válida a capacidade dos dados em representar quadros clínicos reais.

4.3 Pré-processamento e Configuração Experimental

Para lidar com a alta incidência de dados ausentes nas séries temporais, adotou-se uma abordagem inspirada na arquitetura MedLens **[2]**. O processo iniciou-se com a seleção de características (feature selection) baseada na contribuição preditiva dos sinais vitais.

Para a imputação de valores faltantes, diferentemente de métodos estatísticos simples (como média ou mediana), utilizou-se uma técnica de Interpolação baseada em Regressão (Regression-based Interpolation). Especificamente, empregou-se o algoritmo Random Forest Regressor, que aprende padrões não-lineares a partir de amostras completas para estimar valores em pontos temporais específicos. As lacunas remanescentes foram preenchidas utilizando técnicas de forward fill, backward fill e imputação por zero, garantindo a continuidade temporal necessária para os classificadores.

Após a imputação, os dados foram normalizados para padronizar a escala dos atributos numéricos, minimizando vieses decorrentes de magnitudes distintas. A divisão dos dados em conjuntos de treinamento (80%) e teste (20%) foi realizada de forma aleatória e estratificada.

4.3.1 Métricas de Avaliação

As métricas de desempenho são cruciais para comparar modelos. Consideramos a precisão, que mede a proporção de previsões corretas. Embora intuitiva, ela pode ser enganosa em dados desequilibrados, favorecendo a classe majoritária. Para mitigar isso, também usamos o F1-score, a média harmônica da precisão (proporção de verdadeiros positivos entre as previsões positivas) e da recuperação (proporção de verdadeiros positivos entre as instâncias positivas).

Um F1-score mais alto indica um melhor equilíbrio entre precisão e recuperação. Além disso, usamos AUC-ROC e AUC-PR. AUC-ROC (área sob a curva característica de operação do receptor) é um gráfico da taxa de verdadeiros positivos em relação à taxa de falsos positivos, em que um valor mais alto indica melhor desempenho na distinção entre classes. AUC-PR (área sob a curva de precisão-recall) mede a precisão das previsões de mortalidade, com um valor mais alto indicando uma redução nos alarmes falsos.

5 RESULTADOS

A avaliação experimental foi conduzida em duas etapas complementares. Inicialmente, analisou-se

o desempenho descritivo dos classificadores propostos, confrontando-os analiticamente com benchmarks da literatura. Em seguida, aplicou-se um protocolo de validação estatística com múltiplas execuções para verificar a robustez e a significância das diferenças observadas entre os modelos.

Tabela 1- Distribuição Percentual da Distribuição das Partes de um Manuscrito*

MODELO	ACURÁCIA	F1-SCORE	AUC-ROC	AUC-PR
LightGBM	0.930	0.927	0.846	0.667
XGBoost	0.928	0.922	0.851	0.672
AdaBoost	0,932	0.923	0.724	0.456
SVC Calibrado	0.932	0.923	0.769	0.453
Kernel Ridge	0,929	0,919	0.608	0.204

Fonte: Os autores.

Os resultados obtidos (Tabela 1) revelam um trade-off importante entre acurácia global e capacidade de discriminação. Ao observar a métrica de Acurácia, os modelos AdaBoost e SVC Calibrado obtiveram os maiores valores nominais (0,932), superando ligeiramente as abordagens baseadas em Gradient Boosting. No entanto, uma análise mais profunda das métricas de sensibilidade revela limitações nessas abordagens.

Apesar da alta acurácia, o SVC e o AdaBoost apresentaram valores de AUC-PR (Área sob a Curva de Precisão-Revocação) de apenas 0,453 e 0,456, respectivamente. Em contraste, o LightGBM e o XGBoost mantiveram acurácias competitivas (0,930 e 0,928), mas atingiram valores de AUC-PR significativamente superiores (0,667 e 0,672) e AUC-ROC acima de 0,84. Isso indica que, embora o SVC e o AdaBoost classifiquem corretamente a maioria dos casos (provavelmente a classe majoritária de sobreviventes), os modelos de Gradient Boosting (LightGBM e XGBoost) são muito mais eficazes em ordenar corretamente os pacientes de risco e distinguir entre óbito e sobrevida.

Consequentemente, para a tarefa crítica de predição de mortalidade, o F1-Score assume o papel de métrica decisiva. Neste quesito, o LightGBM liderou com 0,927, seguido pelo SVC e AdaBoost (0,923). A combinação de maior F1-Score com as melhores métricas de curva (AUC) consolida o LightGBM e o XGBoost como as abordagens mais robustas e clinicamente úteis.

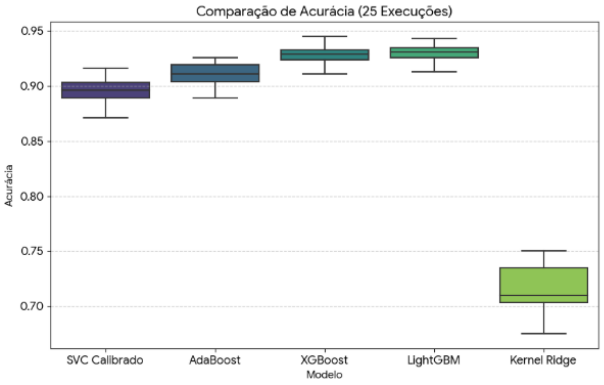
Comparativamente à literatura, todos os modelos propostos superaram os resultados de Harutyunyan et al. **[1]**, onde o KNN e MLP atingiram F1-Scores de 0,869 e 0,848. O Kernel Ridge (KR), embora tenha apresentado acurácia elevada (0,929), obteve a pior capacidade de discriminação (AUC-PR 0,204), sugerindo um overfitting na classe majoritária.

5.1 Validação Estatística e Teste de Hipótese

Para determinar a superioridade estatística entre os modelos líderes em F1-Score e AUC (LightGBM e XGBoost), realizou-se uma validação rigorosa baseada em 25 execuções independentes. O protocolo experimental iniciou-se com a aplicação do teste de Shapiro-Wilk nas distribuições de desempenho das 25 rodadas, onde todos os modelos apresentaram valores-p superiores a 0,05, indicando aderência à distribuição normal e validando o uso de testes paramétricos.

Adicionalmente, para corroborar a robustez dos achados e mitigar eventuais violações de pressupostos, aplicou-se o teste não-paramétrico de Wilcoxon Signed-Rank. O resultado confirmou a significância estatística ($p < 0,05$), demonstrando que a superioridade do modelo proposto independe da distribuição subjacente dos dados.

Figura 3 – Análise Visual da Dispersão dos Classificadores



Fonte: Autores.

DOI: 10.25286/repa.v11i3.3974

A análise visual da dispersão apresentada na Figura 3 revela nuances importantes sobre o comportamento dos classificadores. Observa-se que o LightGBM e o XGBoost se destacam não apenas pelas medianas de acurácia mais elevadas — situadas no topo da escala visual —, mas também pela consistência de seus resultados.

Especificamente, o boxplot do LightGBM apresenta uma amplitude interquartílica (IQR) reduzida (a altura da "caixa"), o que indica uma baixa variabilidade nas predições entre as 25 execuções independentes. Isso sugere que o modelo é altamente estável e pouco sensível às perturbações estocásticas na amostragem dos dados de treinamento. Em contraste, modelos como o SVC Calibrado e o Kernel Ridge demonstram não apenas um desempenho mediano inferior, mas também uma dispersão maior (caixas mais alongadas) e distantes dos líderes.

Embora a sobreposição entre as caixas do LightGBM e do XGBoost sugira uma competitividade acirrada, a concentração dos dados do LightGBM em valores superiores, com menor incidência de outliers negativos, reforça a conclusão estatística anterior: trata-se da abordagem mais eficaz e robusta para este cenário clínico, oferecendo maior garantia de reprodutibilidade em aplicações práticas.

6 CONCLUSÃO

Este trabalho cumpriu o objetivo de aplicar e validar estatisticamente algoritmos de Machine Learning para a predição de mortalidade em unidades de terapia intensiva, utilizando um subconjunto representativo da base MIMIC-III. A estratégia de pré-processamento, fundamentada na metodologia MedLens e aprimorada por técnicas de interpolação, mostrou-se eficaz para mitigar a alta prevalência de dados ausentes e o desbalanceamento de classes, desafios críticos na modelagem de dados de saúde.

Os resultados evidenciaram que os modelos de ensemble superaram as abordagens tradicionais, com o LightGBM consolidando-se como a arquitetura mais robusta, alcançando acurácia média de 92,99% e F1-Score de 0,927. A superioridade técnica do LightGBM sobre concorrentes como o XGBoost pode ser atribuída à sua estratégia de crescimento de árvores baseada em folhas (leaf-wise), combinada com o uso de Amostragem Unilateral Baseada em Gradiente (GOSS). Essas características permitiram que o modelo lidasse de forma mais eficiente com a esparsidade dos dados clínicos e capturasse padrões não-lineares complexos nos sinais vitais, priorizando instâncias com maiores erros de gradiente durante o treinamento.

Mais do que métricas pontuais, a principal contribuição científica deste estudo reside no rigor metodológico. A execução de um protocolo de validação com 25 repetições independentes, seguida pela aplicação do Teste t Pareado, confirmou com 95% de confiança ($p < 0,05$) que o desempenho do LightGBM é estatisticamente significativo e não fruto de aleatoriedade estocástica.

Do ponto de vista da aplicação prática, os resultados sugerem que o modelo proposto é viável para integração em Sistemas de Suporte à Decisão Clínica (CDSS) em tempo real. Diferentemente de escores estáticos tradicionais, a solução baseada em LightGBM oferece alta velocidade de inferência e baixo consumo de memória, permitindo o processamento contínuo de sinais vitais à beira do leito. A implantação desta ferramenta em ambiente hospitalar tem o potencial de automatizar a estratificação de risco, reduzindo a carga cognitiva da equipe médica e permitindo uma alocação de recursos (como leitos de UTI e ventilação mecânica) mais ágil e baseada em evidências, impactando diretamente na redução da mortalidade e na eficiência operacional da unidade.

6.1 Limitações e Trabalhos Futuros

Apesar dos avanços demonstrados, este estudo apresenta limitações inerentes que delineiam oportunidades para investigações subsequentes. A principal restrição refere-se ao uso de um subconjunto estratificado de 30% da base MIMIC-III. Embora essa amostragem tenha sido validada estatisticamente quanto à sua representatividade e eficiência, a utilização da base integral poderia capturar variações inter-individuais mais raras e oferecer um volume de treinamento ainda maior. Como direção primordial para a continuidade desta pesquisa, sugere-se uma comparação direta e exaustiva entre a abordagem aqui desenvolvida e a implementação original completa do sistema MedLens, utilizando a totalidade dos dados do MIMIC-III. Tal experimento permitirá quantificar

com precisão o ganho marginal real da estratégia de interpolação proposta frente ao pipeline completo original.

Em paralelo, recomenda-se a expansão do escopo técnico e geográfico da validação. No âmbito algorítmico, a investigação de arquiteturas de Deep Learning capazes de extrair automaticamente características temporais complexas, como redes Long Short-Term Memory (LSTM) ou Transformers, apresenta-se como um caminho promissor para complementar a eficácia dos modelos de árvore aqui validados. Já no contexto de aplicabilidade clínica, é fundamental avaliar a transferibilidade dos modelos para outros cenários por meio de validação externa em repositórios públicos distintos, como o DATASUS, ou em bases locais de hospitais públicos de Pernambuco. Por fim, a exploração de técnicas mais sofisticadas de engenharia de atributos poderá melhorar a interpretabilidade clínica dos modelos, facilitando sua adoção como ferramentas de "caixa branca" por equipes médicas.

REFERÊNCIAS

[1] HARUTYUNYAN, H.; KHACHATRIAN, H.; KALE, D. C.; VER STEEG, G.; GALSTYAN, A. Multitask learning and benchmarking with clinical time series data. Scientific Data, [S.l.], v. 6, n. 1, p. 96, 2019.

[2] WANG, J. et al. MedLens: Improve Mortality Prediction via Medical Signs Selecting and Regression. **IEEE Journal of Biomedical and Health Informatics**, [S.l.], v. 25, n. 2, p. 390–400, 2021.

[3] JOHNSON, A. E. W. et al. MIMIC-III, a freely accessible critical care database. **Scientific Data**, [S.l.], v. 3, n. 160035, 2016.

[4] CHEN, Tianqi; GUESTRIN, Carlos. **XGBoost: A Scalable Tree Boosting System**. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016. p. 785–794.

[5] KE, Guolin et al. **LightGBM: A Highly Efficient Gradient Boosting Decision Tree**. In: Advances in Neural Information Processing Systems (NIPS 30). Long Beach, CA: Curran Associates, Inc., 2017. p. 3146–3154.

[6] FREUND, Yoav; SCHAPIRE, Robert E. **A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting**. Journal of Computer and System Sciences, v. 55, n. 1, p. 119–139, 1997.

[7] CORTES, Corinna; VAPNIK, Vladimir. **Support-vector networks**. Machine Learning, v. 20, n. 3, p. 273–297, 1995.

[8] PLATT, John. **Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods**. In: Advances in Large Margin Classifiers. Cambridge, MA: MIT Press, 1999. p. 61–74.

[9] SAUNDERS, Craig; GAMMERAAN, Alexander; VOLKMER, Vovk. **Ridge Regression Learning Algorithm in Dual Variables**. In: Proceedings of the 15th International Conference on Machine Learning (ICML). San Francisco: Morgan Kaufmann, 1998. p. 515–521.

[10] HU, T. L. et al. Machine Learning-Based Predictions of Mortality and Readmission in Type 2 Diabetes Patients in the ICU. **Applied Sciences**, [S.l.], v. 14, n. 18, p. 8443, 2024.

[11] IM, S.; LEE, S. M. Development of mortality prediction model using electronic health record (EHR) data and machine learning algorithm in intensive care unit (ICU). **Journal of the Korean Data Analysis Society**, [S.l.], v. 25, n. 5, p. 1977–1994, 2023.

[12] JAIN, E.; SINGH, A. Optimizing Gradient Boosting Algorithms for Obesity Risk Prediction. In: **IEEE INTERNATIONAL CONFERENCE ON CYBERNETICS AND COMPUTATIONAL INTELLIGENCE (CYBERCOM)**, 2024. Proceedings... [S.l.]: IEEE, 2024.

[13] MONTGOMERY, Douglas C. **Design and Analysis of Experiments**. 8. ed. New York: John Wiley & Sons, 2012.

[14] SCHMIDT, G. A.; HALL, J. B. Hypoxemia. In: HALL, J. B.; SCHMIDT, G. A.; KRESS, J. P. (Ed.). Principles of Critical Care. [S.l.]: **McGraw-Hill Education**, 2011. p. 144-154.

DOI: 10.25286/repa.v11i3.3974