ANÁLISE COMPARATIVA DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA CLASSIFICAÇÃO DO RISCO DE SARCOPENIA EM IDOSOS NÃO INSTITUCIONALIZADOS

João Víctor Rodrigues[1] - jvr2564@gmail.com
Alfredo Maurício Batista de Paula[1] - alfredo.paula@unimontes.br
Luciana Mara Barbosa Pereira[1] - lmbp.odonto@gmail.com
Cleiane Gonçalves de Oliveira[2] - cleiane.oliveira@ifnmg.edu.br
[1]Universidade Estadual de Montes Claros, UNIMONTES - Montes Claros, MG, Brazil
[2]Instituto Federal do Norte de Minas Gerais, IFNMG - Januária, MG, Brazil

Abstract. *O rastreamento do risco de sarcopenia pode favorecer a identificação de pessoas idosas que necessitam de avaliação clínica detalhada. Este estudo comparou cinco algoritmos de aprendizagem de máquina na classificação do risco de sarcopenia, definido pela variável SARC-F, em 708 idosos não institucionalizados. O fluxo analítico foi organizado segundo o CRISP-DM. Após a exclusão de identificadores, variáveis com elevada ausência de dados, medidas redundantes ou derivadas e preditores incompatíveis com o cenário proposto, permaneceram 77 preditores. A base foi dividida de forma estratificada em treinamento (80%) e teste (20%). Regressão logística, Naive Bayes, Random Forest, Gradient Boosting e LightGBM foram comparados mediante validação cruzada estratificada com dez partições, sem busca de hiperparâmetros. A comparação considerou métricas tradicionais para classificadores, tais como acurácia, precisão, sensibilidade, especificidade, F1-score e área sob a curva ROC (AUC). O LightGBM apresentou o maior F1 médio na validação cruzada, com desempenho próximo ao Gradient Boosting, e manteve resultados consistentes na base de teste. Os resultados mostram a utilidade de uma comparação padronizada entre diferentes estratégias de aprendizagem e oferecem uma base metodológica para estudos futuros de validação e aprimoramento dos modelos.*

Keywords: *Aprendizado de máquina, Sarcopenia, Classificação, Modelos Computacionais*

1. INTRODUÇÃO

Segundo estimativas da Organização Mundial da Saúde, uma em cada seis pessoas será idosa nas Américas em 2030 (OMS, 2022). Com o envelhecimento populacional, ocorre o aumento da prevalência de condições patológicas cardiovasculares, metabólicas e inflamatórias

crônicas, como a sarcopenia que se configura como uma síndrome geriátrica caracterizada pela perda progressiva da força, da massa e do desempenho muscular esquelético (Barry et al., 2024; Cruz-Jentoft et al., 2019).

A origem da sarcopenia é multifatorial e envolve modificações musculares relacionadas a processos metabólicos e inflamatórios, com possíveis consequências para a capacidade funcional e a qualidade de vida da pessoa idosa (Jamshidi et al., 2022). No presente trabalho, o SARC-F é utilizado como instrumento de rastreamento, portanto, seu resultado indica risco e não substitui o diagnóstico clínico.

Técnicas de aprendizagem de máquina podem explorar simultaneamente diferentes características clínicas, funcionais, antropométricas e sociodemográficas, identificando padrões em conjuntos de dados complexos (França et al., 2021). Contudo, uma comparação útil requer que os modelos sejam avaliados sob as mesmas condições de preparação dos dados, divisão da amostra e validação. Modelos com acurácia semelhante podem apresentar comportamentos distintos na identificação da classe de risco, especialmente quando as classes possuem proporções diferentes. Por isso, a análise deve considerar também métricas como sensibilidade, precisão, F1-score e AUC.

Assim, este estudo teve como objetivo comparar o desempenho de cinco algoritmos de aprendizagem de máquina na classificação do risco de sarcopenia segundo o SARC-F em idosos não institucionalizados. Para isso, os modelos foram submetidos ao mesmo pré-processamento e procedimento de validação, e o modelo selecionado foi posteriormente avaliado em um conjunto de teste preservado durante o treinamento. O estudo apresenta uma comparação computacional, sem pretensão de validar uma ferramenta diagnóstica para uso clínico.

2. METODOLOGIA

2.1 Delineamento e fonte dos dados

Trata-se de uma análise secundária, transversal, observacional e quantitativa de dados previamente coletados com idosos não institucionalizados atendidos no Centro Mais Vida de Referência em Assistência à Saúde do Idoso (CRASI), vinculado ao Hospital Universitário Clemente de Faria da Universidade Estadual de Montes Claros. A base analisada contém 708 participantes e 105 variáveis, abrangendo características sociodemográficas, hábitos de vida, condições clínicas, uso de medicamentos, medidas antropométricas, estado nutricional, saúde bucal e desempenho físico.

2.2 Metodologia CRISP-DM

O fluxo foi organizado segundo o CRISP-DM (*Cross-Industry Standard Process for Data Mining*), modelo de referência composto pelas etapas de compreensão do problema, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação (Wirth & Hipp, 2000). A escolha fornece uma sequência clara para relacionar a pergunta do estudo às decisões de tratamento e avaliação dos dados.

2.3 Preparação dos dados

A variável-alvo foi SARCF, codificada em duas classes: sem risco (508 participantes; 71,8%) e com risco (200 participantes; 28,2%). Assim, os modelos estimaram a classe de

risco indicada pelo instrumento, e não o diagnóstico clínico de sarcopenia.

Foram excluídos três identificadores (COD, NUM e NOME), três variáveis com pelo menos 40% de valores ausentes (RENDA, ANOREXIA e GOHAI), dezesseis medidas derivadas ou redundantes e cinco variáveis que representavam outros desfechos ou informações excessivamente próximas ao fenômeno e à definição do alvo (FRIED, FRAGILIDADE, SARCOPENIA_SDOC, SARCOPENIA_EWGSOP2 e QUEDAS). Ao final dessa etapa, o conjunto reuniu 77 preditores, dos quais 25 eram categóricos e 52 numéricos.

Após as exclusões, o conjunto de dados não apresentava valores ausentes. O pré-processamento foi aplicado de forma idêntica a todos os algoritmos: as variáveis categóricas foram codificadas pelo fluxo de modelagem e as numéricas foram normalizadas pelo escore-z. Para evitar vazamento de informação, os parâmetros dessas transformações foram estimados somente com os dados de treinamento de cada etapa e depois aplicados aos dados correspondentes de validação ou teste.

2.4 Modelagem

As análises foram implementadas na linguagem Python, utilizando a biblioteca PyCaret e bibliotecas associadas de ciência de dados e visualização. A amostra foi dividida uma única vez e de forma estratificada pela variável-alvo, sendo destinados 80% dos participantes ao conjunto de treinamento (566 participantes) e 20% ao conjunto de teste (142 participantes). O conjunto de teste permaneceu separado durante o treinamento e a comparação inicial. No conjunto de treinamento, todos os modelos foram avaliados por validação cruzada estratificada com dez partições e embaralhamento, utilizando as mesmas subdivisões da amostra.

Foram comparados cinco algoritmos representativos de diferentes estratégias de aprendizagem: regressão logística, como método linear e interpretável; Naive Bayes, como método probabilístico; Random Forest, como *ensemble* de árvores baseado em *bagging*; e Gradient Boosting Classifier e LightGBM, como *ensembles* de árvores baseados em *boosting*.

2.5 Avaliação dos modelos

Na validação cruzada foram calculadas a média e o desvio-padrão de acurácia, AUC, sensibilidade (*recall*), precisão, F1-score e coeficiente kappa. A tabela comparativa foi ordenada por acurácia apenas para facilitar a leitura, sem tratá-la como critério único de qualidade. O modelo foi selecionado pelo maior F1 médio na validação cruzada, métrica correspondente à média harmônica entre precisão e sensibilidade.

Depois da seleção, o modelo com maior F1 médio foi aplicado ao conjunto de teste. Foram analisadas acurácia, precisão, sensibilidade, especificidade, F1-score e AUC, juntamente com a matriz de confusão. Uma curva ROC única reuniu os cinco modelos no conjunto de teste para facilitar a comparação visual de sua capacidade discriminativa. Essa figura teve finalidade descritiva e não foi usada para refazer a seleção.

3. RESULTADOS E DISCUSSÃO

3.1 Comparação dos modelos na validação cruzada

A Tabela 1 apresenta a média e o desvio-padrão das métricas obtidas nas dez partições da validação cruzada. A tabela foi organizada em ordem decrescente de acurácia, embora a seleção

do modelo tenha sido definida pelo maior F1 médio.

Table 1: Desempenho dos modelos na validação cruzada estratificada

Modelo	Acurácia	AUC ROC	Sensibilidade	Precisão	F1
LightGBM	$0{,}920 \pm 0{,}033$	$0{,}965 \pm 0{,}020$	$0{,}912 \pm 0{,}064$	$0{,}830 \pm 0{,}072$	$0{,}867 \pm 0{,}053$
GBC	$0{,}918 \pm 0{,}029$	$0{,}961 \pm 0{,}030$	$0{,}887 \pm 0{,}054$	$0{,}837 \pm 0{,}053$	$0{,}861 \pm 0{,}049$
RF	$0{,}871 \pm 0{,}052$	$0{,}937 \pm 0{,}040$	$0{,}725 \pm 0{,}129$	$0{,}803 \pm 0{,}109$	$0{,}757 \pm 0{,}108$
RL	$0{,}846 \pm 0{,}047$	$0{,}898 \pm 0{,}037$	$0{,}700 \pm 0{,}136$	$0{,}746 \pm 0{,}087$	$0{,}716 \pm 0{,}097$
NB	$0{,}733 \pm 0{,}065$	$0{,}814 \pm 0{,}030$	$0{,}700 \pm 0{,}054$	$0{,}535 \pm 0{,}066$	$0{,}601 \pm 0{,}038$

Conforme o critério estabelecido, o LightGBM foi selecionado por apresentar o maior F1 médio na validação cruzada, de $0{,}867 \pm 0{,}053$.

3.2 Avaliação no conjunto de teste

No conjunto de teste, o LightGBM apresentou acurácia de 0,887, acurácia balanceada de 0,884, precisão de 0,761, sensibilidade de 0,875, especificidade de 0,892, F1 de 0,814, AUC ROC de 0,942 e AUC precisão-revocação de 0,869.

A Fig. 1 apresenta a matriz de confusão, que registrou 91 verdadeiros negativos, 11 falsos positivos, cinco falsos negativos e 35 verdadeiros positivos. Assim, o modelo classificou corretamente 126 dos 142 participantes. Entre os 40 participantes com risco, 35 foram identificados corretamente; entre os 102 participantes sem risco, 91 foram classificados corretamente.

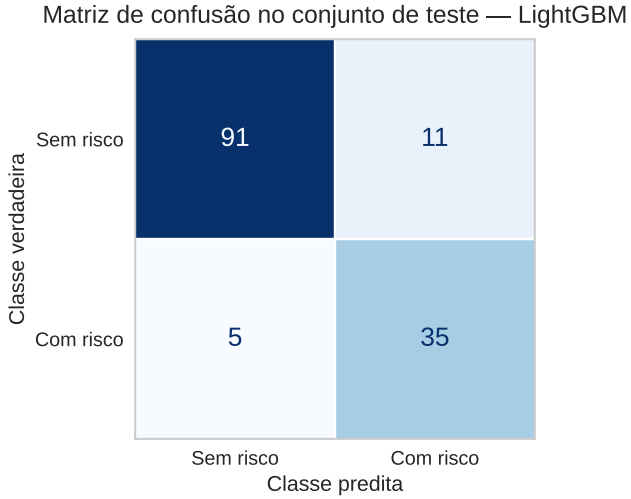

Figure 1: Matriz de confusão no conjunto de teste.

A Fig. 2 reúne as curvas ROC dos cinco modelos no conjunto de teste. As AUCs foram de 0,947 para Random Forest, 0,942 para LightGBM, 0,927 para Gradient Boosting, 0,906 para regressão logística e 0,854 para Naive Bayes. Embora o Random Forest tenha apresentado a maior AUC nesse conjunto, a seleção do LightGBM não foi alterada, pois havia sido realizada anteriormente pelo F1 médio da validação cruzada.

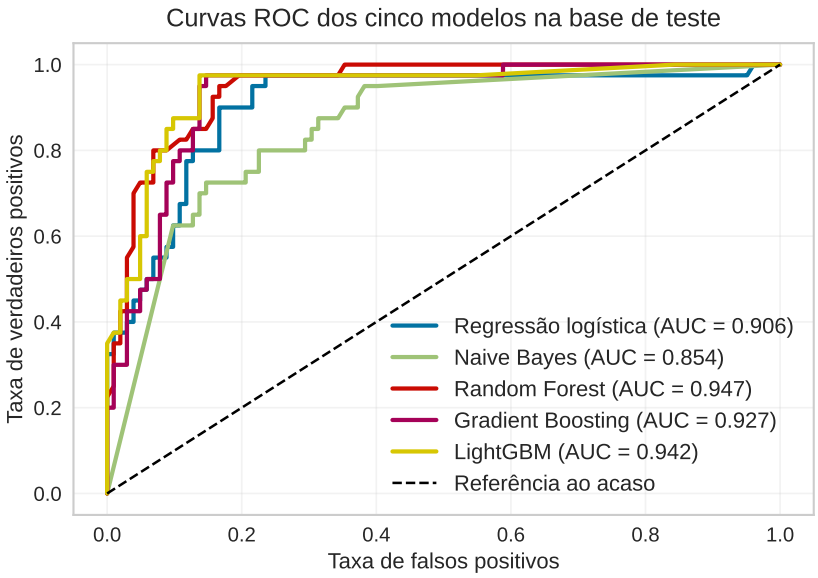

Figure 2: Curvas ROC dos cinco modelos no conjunto de teste.

Os resultados da validação cruzada indicaram melhor desempenho médio dos dois métodos baseados em *boosting*. O LightGBM apresentou o maior F1 médio e foi selecionado conforme o critério previamente definido, enquanto o Gradient Boosting Classifier apresentou valores próximos e a maior precisão. As diferenças entre esses modelos foram inferiores aos respectivos desvios-padrão, não sustentando uma afirmação de superioridade ampla do LightGBM. A principal conclusão da comparação é, portanto, que os dois métodos formaram o grupo de melhor desempenho, com pequena vantagem numérica do LightGBM nas métricas utilizadas para a seleção.

O Random Forest apresentou comportamento distinto conforme a métrica analisada. Embora seu F1 e sua sensibilidade na validação cruzada tenham sido inferiores aos dos métodos de *boosting*, o modelo alcançou a maior AUC ROC no conjunto de teste. Essa mudança de ordenação não invalida a seleção do LightGBM, pois a escolha foi realizada anteriormente pelo F1 médio da validação cruzada. A AUC avalia a capacidade de discriminação em diferentes pontos de corte, enquanto o F1 resume o equilíbrio entre precisão e sensibilidade em um ponto específico de classificação. Dessa forma, modelos diferentes podem se destacar segundo métricas distintas.

A regressão logística apresentou desempenho inferior ao dos métodos baseados em árvores, mas permaneceu como uma referência linear de menor complexidade e interpretação mais direta. O Naive Bayes obteve o menor desempenho geral, principalmente em precisão e F1, indicando maior proporção de classificações positivas incorretas. Esses resultados reforçam que a acurácia não deve ser interpretada isoladamente, sobretudo diante da menor frequência de participantes classificados com risco.

No conjunto de teste, o LightGBM manteve sensibilidade e especificidade próximas, identificando 35 dos 40 participantes com risco e 91 dos 102 participantes sem risco. Em uma possível aplicação de rastreamento, os cinco falsos negativos mereceriam atenção por representarem participantes com risco classificados como sem risco. Entretanto, este estudo realiza uma comparação computacional e não avalia consequências clínicas, custos de erros ou pontos de corte adequados para decisões em saúde.

Entre as limitações, o conjunto de teste continha apenas 40 participantes na classe com risco. Os modelos também foram avaliados com configurações padrão, sem ajuste de hiperparâmetros, análise de calibração ou validação externa. Estudos futuros podem avaliar outras amostras, conjuntos menores de preditores disponíveis em situações reais de rastreamento, métodos de interpretação das variáveis, calibração das probabilidades e ajuste dos modelos. Técnicas de balanceamento e busca de hiperparâmetros podem ser investigadas em experimentos separados, sem comprometer a comparabilidade da análise apresentada.

4. CONCLUSÕES

A comparação mostrou que os algoritmos apresentaram desempenhos distintos na classificação do risco de sarcopenia segundo o SARC-F. LightGBM e Gradient Boosting Classifier formaram o grupo de melhor desempenho na validação cruzada, com pequena vantagem numérica do LightGBM, selecionado pelo maior F1 médio. Na base de teste, o modelo manteve resultados consistentes e equilíbrio entre sensibilidade e especificidade, enquanto o Random Forest apresentou a maior AUC ROC nessa avaliação.

Os achados demonstram a importância de considerar métricas complementares e a variabilidade entre as partições, em vez de selecionar um modelo apenas pela acurácia. Os resultados são preliminares e não representam validação para uso clínico. A continuidade do estudo requer validação externa, avaliação de conjuntos de atributos mais simples e investigação do comportamento dos modelos em diferentes populações e cenários de rastreamento.

Agradecimentos

Os autores agradecem ao CNPq (Processos nº 441306/2023 e 300935/2025-0), à CAPES e à FAPEMIG (grant n OET-00243-26 e Processos RED-031-21, APQ-04800-24 e APQ-08594-25).

REFERÊNCIAS

Barry, D.J. et al. (2024), "Investigating the Effects of Symbiotic Supplementation on Functional Movement, Strength and Muscle Health in Older Australians: A Study Protocol for a Double-Blind, Randomized, Placebo-Controlled Trial", *Trials*, 25(1), 307.

Cruz-Jentoft, A.J. et al. (2019), "Sarcopenia: Revised European Consensus on Definition and Diagnosis", *Age and Ageing*, 48(1), 16–31. doi: 10.1093/ageing/afy169.

França, R.P. et al. (2021), "An Overview of Deep Learning in Big Data, Image, and Signal Processing in the Modern Digital Age", *Trends in Deep Learning Methodologies*, 63–87.

Jamshidi, S. et al. (2022), "The Effects of Symbiotic and/or Vitamin D Supplementation on Gut-Muscle Axis in Overweight and Obese Women: A Study Protocol for a Double-Blind, Randomized, Placebo-Controlled Trial", *Trials*, 23(1), 631.

Organização Mundial da Saúde (2022), *Enfrentando o abuso de idosos: cinco prioridades para a Década das Nações Unidas do Envelhecimento Saudável (2021–2030)*, Organização Mundial da Saúde, Genebra.

Wirth, R.; Hipp, J. (2000), "CRISP-DM: Towards a Standard Process Model for Data Mining", *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29–39. Disponível em: `https://www.cs.unibo.it/~montesi/CBD/Beatriz/10.1.1.198.5133.pdf`. Acesso em: 22 jul. 2026.