IMPORTÂNCIA DE VARIÁVEIS EM MODELOS PREDITIVOS: ABORDAGENS BASEADAS EM REGRESSÃO E INTELIGÊNCIA COMPUTACIONAL

Maria Laucinéia Carari[1], Wagner Faria Barbosa[2], Ana Carolina Campana Nascimento[3], Moysés Nascimento[3], Camila Ferreira Azevedo[3], Gabriela França Oliveira[4], Paulo Roberto Cecon[3]

[1] Doutoranda em Estatística Aplicada e Biometria pela Universidade Federal de Viçosa, Viçosa-MG. E-mail:neiacarari@gmail.com
[2]Pós-doutorando na Empresa de Pesquisa Agropecuária de Minas Gerais, Viçosa-MG
[3] Professores do Departamento de Estatística da Universidade Federal de Viçosa, Viçosa-MG
[4] Pós-doutoranda do Departamento de Estatística da Universidade Federal de Viçosa, Viçosa-MG

Recebido em: 15/05/2025 – Aprovado em: 15/06/2025 – Publicado em: 30/06/2025
DOI: 10.18677/EnciBio_2025B

RESUMO

A compreensão da importância das variáveis explicativas configura-se como uma etapa essencial no ajuste de modelos de regressão e de inteligência computacional (IC). Variáveis de pouca relevância tendem a elevar a complexidade do modelo sem, contudo, promover uma melhoria em sua capacidade preditiva. Nesse contexto, diversas metodologias têm sido desenvolvidas com o objetivo de quantificar a importância das variáveis e, consequentemente, auxiliar na identificação daquelas de maior influência. No entanto, não há consenso na literatura quanto ao método mais eficaz, especialmente considerando diferentes contextos, tamanhos amostrais e objetivos analíticos. Deste modo, este estudo teve como objetivo avaliar cinco metodologias, *Stepwise*, *MARS*, *Bagging*, Garson e Olden, para a determinação da importância de variáveis, utilizando dados simulados. Foram considerados dois cenários, um modelo aditivo com apenas efeitos principais e outro com interações entre variáveis explicativas. As metodologias foram avaliadas com base na taxa de acerto na identificação das variáveis mais importantes, na capacidade de detectar interações e na qualidade preditiva, por meio da capacidade preditiva (CP) e da raiz do erro quadrático médio (REQM). Os resultados mostraram que, em ambos os cenários, os métodos apresentaram desempenho semelhante. O *MARS* destacou-se por sua habilidade em capturar interações de forma explícita e manter bom desempenho preditivo. As redes neurais, avaliadas pelos métodos de Garson e Olden, também apresentaram elevada CP, especialmente em amostras maiores. Os resultados reforçam a necessidade de alinhar a escolha metodológica aos objetivos do estudo, à complexidade dos dados e ao tamanho amostral.
PALAVRAS-CHAVE: Predição, Redes Neurais, Simulação.

VARIABLE IMPORTANCE IN PREDICTIVE MODELS: REGRESSION-BASED AND COMPUTATIONAL INTELLIGENCE APPROACHES

ABSTRACT

Understanding the importance of explanatory variables is an essential step in fitting regression and computational intelligence (CI) models. Irrelevant variables tend to increase model complexity without improving its predictive performance. In this

context, several methodologies have been proposed to quantify the importance of variables and assist in identifying those with the greatest influence. However, there is no consensus in the literature regarding the most effective method, especially when considering different contexts, sample sizes, and analytical purposes. This study aimed to evaluate five methodologies, Stepwise, MARS, Bagging, Garson, and Olden, for determining variable importance using simulated data. Two scenarios were considered: one additive model with only main effects, and another including interactions among explanatory variables. The methods were assessed based on their accuracy in identifying the most important variables, their ability to detect interactions, and their predictive performance, through predictive capacity (PC) and root mean squared error (RMSE). The results showed that all methods performed similarly in both scenarios. MARS stood out for its ability to explicitly capture interactions while maintaining good predictive performance. Neural networks, evaluated through the Garson and Olden methods, also demonstrated high predictive capacity, especially with larger samples. The findings reinforce the importance of aligning methodological choices with study objectives, data complexity, and sample size.
KEYWORDS: Prediction, Neural Networks, Simulation.

INTRODUÇÃO

A identificação das variáveis mais importantes na explicação de um determinado fenômeno de interesse é uma etapa fundamental para a construção de modelos explicativos e preditivos robustos, especialmente em contextos aplicados como as ciências agrárias. A identificação adequada das variáveis explicativas relevantes permite a construção de modelos mais parcimoniosos, com melhor desempenho preditivo, menor complexidade computacional, maior capacidade de generalização e interpretabilidade (HASTIE, *et al*., 2009; CHENG, 2024).

Dentre os métodos baseados em regressão, destacam-se o *Stepwise* e o *MARS* (*Multivariate Adaptive Regression Splines*), amplamente utilizados, pois permitem identificar as variáveis mais importantes, selecionar variáveis e interpretar seus efeitos. O *Stepwise*, método sequencial que adiciona ou remove variáveis em um modelo de regressão com base em critérios estatísticos, como AIC, é mais útil em situações com poucos preditores e relações predominantemente lineares (JAMES *et al.,* 2021). Já o *MARS* é mais flexível, pois modela também, relações não lineares por meio de funções do tipo *spline*, sendo capaz de identificar interações entre variáveis mesmo em modelos complexos (FRIEDMAN, 1991; CELERI, 2024).

Por outro lado, os métodos baseados em inteligência computacional têm ganhado destaque pela capacidade de lidar tanto com estruturas simples, como as mais complexas. O *Bagging* (*Bootstrap Aggregating*) utiliza a média de diversos modelos ajustados a subconjuntos aleatórios dos dados, e a importância das variáveis pode ser estimada com base na variação do erro ao permutar os preditores, por meio da métrica IncMSE (*Increase in Mean Squared Error)* (BREIMAN, 1996; BREIMAN *et al*., 2024). Esse procedimento tem se mostrado eficaz para reduzir a variância de modelos instáveis, como as árvores de decisão.

Já nas redes neurais artificiais, a interpretação da importância das variáveis não é trivial, dado seu caráter de "caixa-preta" (OLDEN; JACKSON, 2002). No entanto, métodos como os propostos por Garson (1991) e Olden *et al.* (2004) buscam atribuir importância às variáveis de entrada com base nos pesos das

conexões entre neurônios. Garson (1991) propôs uma decomposição dos pesos da rede, enquanto Olden *et al.* (2004) introduziram um método que considera tanto magnitude quanto sinal dos pesos, o que melhora a interpretação do efeito de cada variável. Essas abordagens têm sido muito utilizadas na análise de redes neurais aplicadas à agricultura, embora ainda apresentem limitações quanto à detecção de interações explícitas (GOH, 1995; GEVREY *et al.,* 2003).

Em ciências agrárias, conhecer as variáveis mais importantes possui implicações práticas diretas, como a predição de produtividade de cultivos, identificação de fatores que afetam o rendimento, avaliação da resistência a doenças ou resposta a práticas de manejo. Estudos recentes evidenciam o potencial de algumas metodologias, como no estudo de Ibrahim *et al*. (2022), que utilizaram o método de Garson para identificar as principais variáveis responsáveis pela dinâmica populacional de pragas agrícolas, o que pode orientar estratégias de monitoramento e controle. Silva Júnior *et al*. (2022), por sua vez, aplicaram os métodos de Bagging e Garson para analisar a importância de características agronômicas e climáticas na previsão da produtividade do arroz irrigado, identificando variáveis como floração número de grãos por panícula e comprimento da panícula entre as mais relevantes, auxiliando na tomada de decisão.

O uso de dados simulados para avaliação de diferentes metodologias é uma prática comum para avaliar a eficácia dos métodos. Essa abordagem permite validar a taxa de acerto na identificação das variáveis realmente relevantes, dado que a estrutura verdadeira é conhecida, ao contrário dos dados empíricos (OLDEN *et al*., 2004).

Apesar da relevância em se conhecer a importância das variáveis explicativas em modelos de regressão, não há consenso na literatura sobre qual metodologia é mais eficaz em diferentes cenários. A escolha do método depende do equilíbrio entre interpretabilidade e desempenho preditivo, além das características específicas dos dados. Por isso, alguns autores recomendam o uso combinado de técnicas ou, pelo menos, a comparação entre diferentes abordagens para uma avaliação mais robusta da importância das variáveis (MOLNAR, 2022).

Diante do exposto, esse estudo teve como objetivo avaliar diferentes métodos de importância de variáveis, a saber: *Stepwise*, *MARS*, *Bagging*, Garson e Olden, utilizando dados simulados. A avaliação dessas metodologias foi realizada com base na capacidade preditiva, na raiz do erro quadrático médio, na ordem de importância das variáveis e na capacidade de identificar as variáveis realmente importantes para os cenários estabelecidos, visando contribuir para a escolha de métodos adequados a diferentes contextos.

MATERIAIS E MÉTODOS

Os dados utilizados neste estudo foram gerados por simulação, considerando dois cenários distintos, denominados M1 e M2. O cenário M1 representa um fenômeno em que a variável resposta Y é influenciada exclusivamente pelos efeitos principais das variáveis explicativas. Já no cenário M2, além dos efeitos principais, considera-se também um termo de interação entre duas variáveis explicativas, refletindo uma estrutura mais complexa do fenômeno em análise. Os modelos geradores dos dados para cada cenário foram definidos conforme as seguintes expressões:

Modelo 1 (M1): $Y_i = 300 + 3{,}5\,X_1 - 1{,}75\,X_2 + 0{,}01\,X_3 + 0\,X_4 + e_i$

Modelo 2 (M2): $Y_i = 300 + 3{,}5\,X_1 - 1{,}75\,X_2 + 0{,}01\,X_3 + 0\,X_4 + 0{,}8\,X_1 X_2 + e_i$

Nesses modelos, os coeficientes atribuídos às variáveis explicativas (X_1, X_2, X_3 e X_4) refletem seus respectivos efeitos sobre a variável resposta Y. O valor de 3,5 para X_1 indica um efeito positivo expressivo, enquanto X_2 tem um impacto negativo de menor magnitude (−1,75), e X_3 exerce um efeito praticamente nulo (0,01). Para a variável X_4, por sua vez, foi propositalmente atribuído um coeficiente nulo (0) em ambos os modelos, com o objetivo de verificar como os diferentes métodos de determinação da importância de variáveis se comportam na presença de uma variável irrelevante. Além disso, no cenário M2 foi incluído um termo de interação entre X_1 e X_2, com coeficiente 0,80, simulando situações em que o efeito conjunto entre variáveis explicativas influencia a resposta.

Para cada cenário (M1 e M2), foram realizadas dez repetições, com três tamanhos amostrais distintos, n = 50, 100 e 500 observações. Isso resultou em 10 conjuntos de dados para cada combinação de cenário e tamanho amostral.

As variáveis explicativas foram geradas a partir de distribuições normais, com os seguintes parâmetros: $X_1 \sim N(100;100)$, $X_2 \sim N(60;36)$, $X_3 \sim N(80;64)$ e $X_4 \sim N(70;196)$. O termo de erro aleatório (e_i) foi gerado assumindo distribuição normal com média zero e variância 25, ou seja, $e_i \sim N(0;25)$.

Metodologias avaliadas

Foram avaliadas cinco metodologias para identificação da importância de variáveis em modelos preditivos: *Stepwise* e *MARS* (baseadas em regressão), *Bagging* (baseado em árvores de decisão), e os métodos de Garson e Olden (baseados em redes neurais artificiais) (GARSON, 1991; FRIEDMAN, 1991; GOH, 1995; OLDEN, 2004; JAMES *et al*., 2021).

Regressão Linear Múltipla

A regressão linear múltipla consiste em uma generalização do modelo de regressão linear simples, que permite modelar a dependência de uma variável resposta Y em função de um conjunto de *p* variáveis explicativas (X_1, X_2, ..., X_p). Essa abordagem possibilita o estudo da influência de *p* variáveis explicativas sobre a variável resposta Y, caso exista uma relação linear entre as variáveis. O modelo geral pode ser expresso como (MONTGOMERY *et al.,* 2021; NASCIMENTO *et al.*, 2024):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + e_i, i = 1,2,3,...,n,$$

em que, Y_i é a i-ésima observação da variável Y; X_{ij} é a i-ésima observação da j-ésima variável, com $j = 1,2,\ldots,p$; n é o número de observações; $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, são os parâmetros do modelo e e_i é o erro aleatório associado a i-ésima observação.

Stepwise

A metodologia *Stepwise* consiste em uma abordagem automatizada que realiza a seleção de variáveis em modelos lineares, por meio dos procedimentos *forward* (adição) e *backward* (remoção) de variáveis explicativas (JAMES *et al*., 2021). Essa metodologia realiza automaticamente o processo de inclusão ou exclusão das variáveis, no entanto, só avalia os termos previamente especificados no modelo.

Neste estudo, a escolha do modelo foi baseada no critério de informação AIC (*Akaike Information Criterion*), em que valores menores indicam melhor ajuste (AKAIKE, 1973). A importância das variáveis foi determinada por meio da métrica de aumento no erro quadrático médio (*Increase in Mean Squared Error – IncMSE*), uma medida que quantifica o aumento no erro quadrático médio de previsão quando se *perturba* ou remove uma variável do modelo. Em termos práticos, o cálculo do *IncMSE* envolve a permutação aleatória dos valores da variável X_j nos dados de teste, quebrando a associação original entre essa variável e a resposta Y, e posteriormente se calcula o novo erro quadrático médio. A diferença entre o novo erro e o erro original indica a importância da variável:

$$IncMSE(X_j) = MSE_{perm}(X_j) - MSE_{orig},$$

em que MSE_{orig} é o erro quadrático médio do modelo sobre o conjunto de teste original e, $MSE_{perm}(X_j)$ é o erro após a permutação aleatória dos valores da variável X_j.

Esse aumento no erro também pode ser expresso em termos relativos, como percentual:

$$\%IncMSE(X_j) = \left(\frac{MSE_{perm}(X_j) - MSE_{orig}}{MSE_{orig}} \right) \times 100$$

Quanto maior o *IncMSE* (ou %*IncMSE*), maior a contribuição da variável para a capacidade preditiva do modelo (LUNDBERG *et al*., 2020; BREIMAN, 2024).

MARS

A metodologia *MARS* (*Multivariate Adaptive Regression Splines*) é uma técnica de regressão que envolve a geração de funções de base, que segmentam os dados ajustando modelos lineares, com ou sem interações, por partes. Cada função de base é formada por *splines* lineares por partes e utiliza nós para indicar os pontos em que a inclinação da função muda (FRIEDMAN, 1991; HUANG *et al*., 2020; CELERI, 2024).

O modelo *MARS* proposto por Friedman (1991) pode ser descrito como:

$$\hat{f}(X) = c_0 + \sum_{i=1}^{M} c_i B_i(X) + \epsilon'$$

em que c_0 corresponde à constante de regressão, $B_i(X)$ corresponde a uma função ou produto de funções de base, c_i é o coeficiente de B_i, M representa o número de funções ou de produtos de funções de base do modelo que é definido automaticamente pelo algoritmo *MARS* e ϵ corresponde ao erro aleatório (FRIEDMAN, 1991; AL-SUDANI *et al.* 2019).

As análises foram conduzidas com o pacote *earth* no R, e a importância das variáveis foi calculada com base no GVC, critério de Validação Cruzada Generalizada *(Generalized Cross-Validation*) (HASTIE *et al*., 2009; MILBORROW, 2018; AL-SUDANI *et al*., 2019). O GCV pode ser expresso por (HASTIE *et al*., 2009 ; NASCIMENTO *et al*., 2024):

$$GCV = \frac{SQR}{N\left(1 - \frac{r + cK}{N}\right)^2}$$

em que *SQR* é a Soma de Quadrado dos Resíduos, *N* é o número de observações, *r* é o número de funções de base linearmente independentes; *K* é o número de nós do modelo e *c* é o fator de penalização (usualmente fixado em 3). Este critério penaliza a complexidade do modelo e auxilia na escolha do número adequado de termos, evitando o sobreajuste (*overfitting*) (HASTIE *et al.*, 2009; NASCIMENTO *et al.*, 2024).

Bagging

A metodologia *Bagging* (*Bootstrap Aggregating*) é uma técnica amplamente utilizada no contexto do aprendizado de máquina, que visa reduzir a variância e aumentar a robustez dos modelos, especialmente modelos baseados em árvores de decisão (BREIMAN, 1996). O procedimento *Bagging*, consiste em gerar vários subconjuntos de dados, usando todas as variáveis, por meio da técnica de amostragem *bootstrap* (com reposição), ajustar um modelo de árvore de decisão para cada subconjunto e, combinar os resultados desses modelos. No caso da regressão, a agregação final é geralmente realizada por meio da média das previsões (BREIMAN, 1996; JAMES *et al.*, 2021; NASCIMENTO *et al.*, 2024).

Além de aprimorar a capacidade preditiva, o *Bagging* também permite estimar a importância das variáveis explicativas, que foi estimada, neste estudo, por meio da métrica *IncMSE* (*Increase in Mean Squared Error*), que determina o aumento no erro de previsão causado pela permutação aleatória dos valores de uma variável (BREIMAN, 1996).

Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs), inspiradas no funcionamento do cérebro humano, têm sido amplamente utilizadas para a modelagem de diversos fenômenos (CRUZ; NASCIMENTO, 2018). Neste estudo, para a estimativa da importância das variáveis por meio dos métodos de Garson e de Olden, foram utilizadas duas arquiteturas distintas de redes do tipo Perceptron Multicamadas (Multilayer Perceptron – MLP), denominadas RN1 e RN2.

As configurações de RN1 e RN2 foram utilizadas para avaliar as metodologias de Garson e Olden. A RN1, aplicada aos métodos de Garson e Olden, implementada com a função nnet() do pacote *nnet*, em R, apresentou uma camada de entrada com quatro variáveis explicativas, uma camada oculta (com 1 a 100 neurônios, em 28 configurações) e uma camada de saída linear. A configuração RN2, exclusiva para o método de Olden, foi implementada com a função mlp() do pacote *RSNNS*, permitindo de uma a três camadas ocultas, com combinações de 2, 5, 10 e 25 neurônios por camada (ex.: (2), (5,10), (10, 25, 2)). Ambas empregaram função de ativação logística, saída linear, pesos inicializados entre [−0,3; 0,3], regularização (decay = 0,005), e até 5.000 iterações. Os dados (normalizados em RN2) foram divididos em 70% para treino e 30% para teste, nos cenários simulados M1 e M2, com amostras de 50, 100 e 500 observações, totalizando 1.680 modelos. A avaliação considerou métricas como correlação, erro quadrático médio (EQM) e raiz do erro quadrático médio (REQM).

Garson

O método de Garson (1991), com aprimoramentos propostos por Goh (1995), baseia-se no particionamento dos pesos de conexão das redes neurais artificiais (RNAs) para estimar a importância relativa de cada variável de entrada. O algoritmo

identifica os pesos que ligam cada variável explicativa à saída, passando pela camada oculta, e calcula sua importância relativa com base na soma e normalização desses pesos. O resultado é um valor absoluto entre 0 e 1, refletindo a magnitude da contribuição de cada variável. No entanto, o método não considera a direção (sinal positivo ou negativo) da influência e é aplicável apenas a redes com uma única camada oculta (GARSON, 1991; GOH, 1995).

Olden

O método de Olden *et al*. (2004) também utiliza os pesos das conexões para estimar a importância das variáveis, mas considera o produto dos pesos brutos entre as camadas, preservando tanto a magnitude quanto o sinal das contribuições. Essa abordagem permite avaliar o impacto positivo ou negativo de cada variável explicativa, podendo ser aplicável a redes com múltiplas camadas ocultas (GEVREY *et al*., 2003; OLDEN *et al*., 2004). Isso torna o método particularmente útil em contextos em que a direção do efeito é relevante, como nas ciências agrárias, em que compreender se o impacto de uma variável é benéfico ou prejudicial pode orientar práticas de manejo mais eficazes.

Avaliação das metodologias

Buscou-se verificar quais metodologias foram eficazes em identificar as variáveis preditoras realmente importantes e em realizar predições para o fenômeno estudado. Para tanto, os modelos foram treinados utilizando 70% dos dados disponíveis enquanto os 30% restantes foram destinados à etapa de teste. O processo foi repetido 10 vezes para garantir uma análise mais robusta e menos suscetível a variações ocasionais, visando garantir maior confiabilidade dos resultados.

As metodologias foram avaliadas com base nos seguintes critérios: capacidade de identificar corretamente as variáveis mais importantes nos dados simulados, capacidade preditiva (CP) dos modelos, mensurada pela correlação entre os valores observados e preditos, na raiz do erro quadrático médio (REQM), que quantifica o desvio médio dos valores preditos em relação aos valores reais e, com relação à similaridade dos métodos na identificação das variáveis mais relevantes nos dois modelos. Esse processo não apenas permitiu verificar o desempenho preditivo dos modelos, mas também avaliar sua capacidade de identificar as variáveis explicativas com impacto significativo na variável resposta.

Aspectos Computacionais

As análises foram realizadas no software R (R CORE TEAM, 2023), versão 4.3.1, com os pacotes *MASS, earth, randomForest, nnet* e *RSNNS*. (BERNARDI; GÜNTHER, 2022; VENABLES; RIPLEY, 2023; BREIMAN, 2024; RIPLEY; VENABLES, 2025; MILBORROW, 2025).

RESULTADOS E DISCUSSÃO

As Figuras 1 e 2 apresentam a frequência com que as variáveis explicativas foram ranqueadas entre as mais importantes por cada uma das metodologias avaliadas, considerando os diferentes tamanhos amostrais e os dois cenários de simulação: M1 (modelo aditivo com apenas efeitos principais) e M2 (modelo aditivo com interação entre variáveis).

De maneira geral, as metodologias avaliadas obtiveram resultados semelhantes na identificação das variáveis mais importantes, X_1 e X_2, independentemente do cenário e do tamanho amostral. Essa consistência foi particularmente evidente no cenário M1, no qual todas as metodologias classificaram, na ordem correta, X_1 e X_2 como as duas variáveis mais importantes nas dez repetições (Figura 1). Com relação às variáveis de menor importância (X_3 e X_4), observou-se maior taxa de acerto na ordenação correta da importância com o aumento do tamanho da amostra para 500 observações, indicando que amostras maiores tendem a produzir resultados mais estáveis e robustos (JAMES *et al*., 2021).

A metodologia *MARS* destacou-se por manter a correta identificação das duas variáveis mais importantes, X_1 e X_2, em ambos os cenários (Figuras 1 e 2), inclusive mantendo sua ordem mesmo na presença de interações no cenário M2 (Figura 2). Isso pode ser atribuído à capacidade da *MARS* de modelar interações entre as variáveis de maneira automática durante o processo de ajuste (FRIEDMAN, 1991), o que representa uma vantagem sobre métodos como o *Stepwise*, que dependem da inclusão manual desses termos de interação para avaliá-los (JAMES *et al*., 2021).

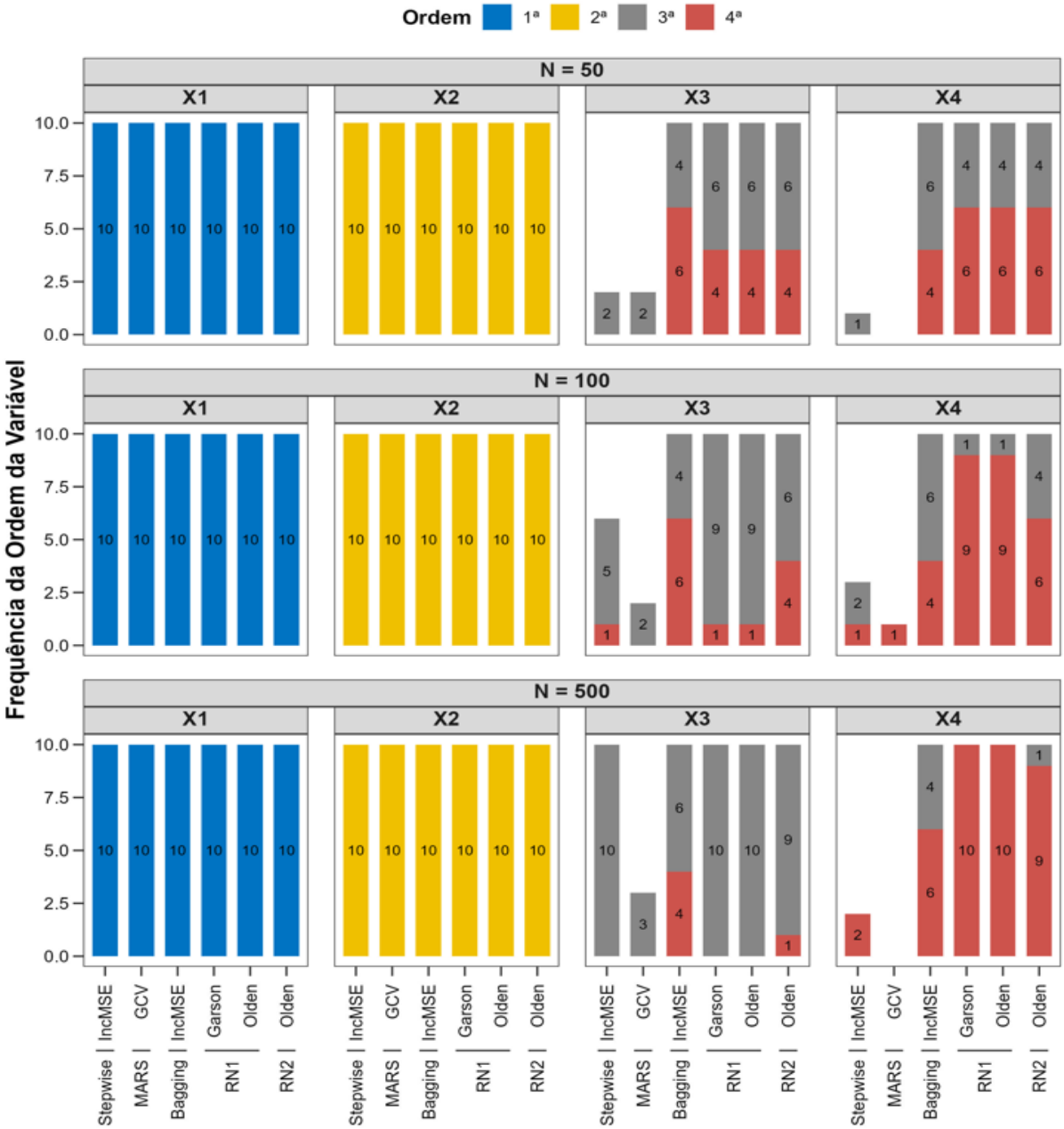
No cenário M2 (Figura 2), observou-se certa variação na ordem de importância entre as variáveis X_1 e X_2 para as metodologias *Stepwise*, *Bagging*, Garson e Olden, com X_2 sendo eventualmente classificada como a mais importante. Essa alteração pode estar relacionada a presença de interação, que pode alterar a contribuição das variáveis explicativas. Ainda assim, essa variação foi menos observada nas amostras maiores.

As metodologias Garson e Olden, aplicadas em redes neurais artificiais, apresentaram resultados semelhantes entre si, com boa capacidade de identificar corretamente as variáveis mais importantes, especialmente nos maiores tamanhos amostrais (Figuras 1 e 2). Embora ambos os métodos utilizem os pesos das conexões das redes para estimar a importância das variáveis, a literatura aponta que o método proposto por Olden (OLDEN; JACKSON, 2002) pode ser preferível ao método de Garson (1991) por considerar não apenas a magnitude dos pesos das conexões, mas também seus sinais, o que permite melhor avaliação sobre a contribuição positiva ou negativa de cada variável (GEVREY *et al*., 2003). Além disso, o método de Olden pode ser aplicado a redes com múltiplas camadas ocultas, enquanto o de Garson é mais restrito a arquiteturas de camada única.

No presente estudo, no entanto, ambos os métodos apresentaram desempenho semelhante, para os dois cenários, possivelmente devido à baixa complexidade dos modelos simulados. Vale destacar que, embora redes neurais sejam capazes de capturar relações não lineares complexas e interações implícitas entre variáveis, os métodos de Garson e Olden não são capazes de separar explicitamente os efeitos de interação entre variáveis explicativas (GEVREY *et al*., 2003; OLDEN, 2004).

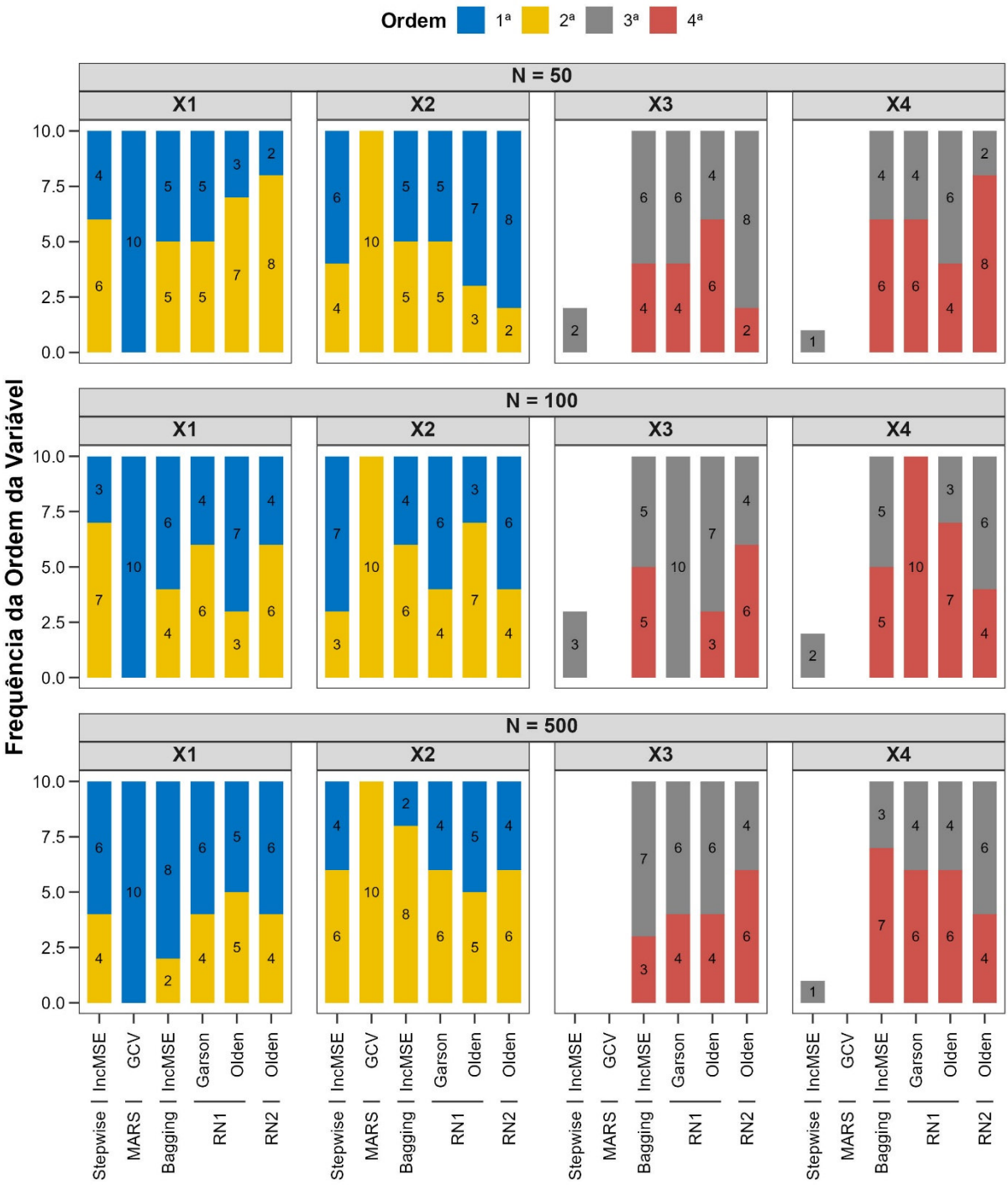
Em relação a variável X_4, incluída nos modelos simulados com coeficiente nulo, a metodologia *MARS* foi a única que identificou a não significância dessa variável em todos os cenários e tamanhos amostrais (Figuras 1 e 2). Esse desempenho está associado à estratégia de seleção baseada no critério GCV, que penaliza variáveis sem contribuição significativa (FRIEDMAN, 1991). Já os métodos Stepwise, Bagging, Garson e Olden atribuíram, em diferentes graus, importância residual à X_4, evidenciando menor capacidade de discriminação frente a variáveis irrelevantes.

FIGURA 1 – Frequência da ordem de importância das variáveis pelas metodologias *Stepwise*, MARS, *Bagging*, Garson e Olden considerando o cenário M1, para n=50, 100 e 500 em 10 repetições.



Fonte: Autores (2025).

FIGURA 2- Frequência da ordem de importância das variáveis pelas metodologias Stepwise, MARS, Bagging, Garson e Olden considerando o cenário M2, para n=50, 100 e 500 em 10 repetições.



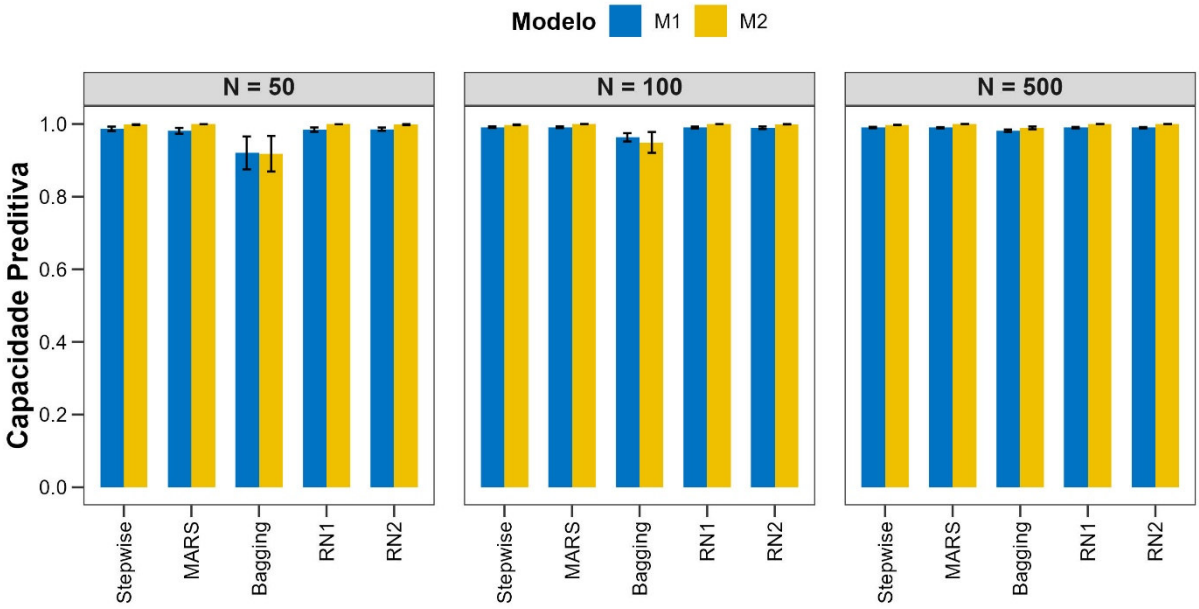
Fonte: Autores (2025).

Além de identificar as variáveis mais importantes, as metodologias também foram avaliadas com base na estimativa média das CP e REQM, com os respectivos EPM, conforme Figuras 3 e 4. As metodologias analisadas mostraram desempenhos

preditivos (CP) semelhantes nos dois cenários, com melhora progressiva associada ao aumento do tamanho amostral (Figuras 3 e 4). O aumento do desempenho preditivo e a redução dos erros associados as amostras maiores estão em concordância com a literatura, que aponta que amostras maiores contribuem para modelos mais estáveis e com menor variabilidade dos erros (JAMES *et al*., 2021).

No cenário M2, foi observada melhora na capacidade preditiva em comparação ao cenário M1 para a maioria das metodologias, com exceção do *Bagging* nas menores amostras (Figura 3). Contudo, os valores de REQM foram, em geral, mais elevados em M2, refletindo a maior complexidade estrutural do modelo (Figura 4).

FIGURA 3 - Estimativa média das capacidades preditivas para as metodologias *Stepwise*, *MARS*, *Bagging* e Redes Neurais, juntamente com os respectivos erros-padrão da média considerando os cenários M1 e M2 em 10 repetições.

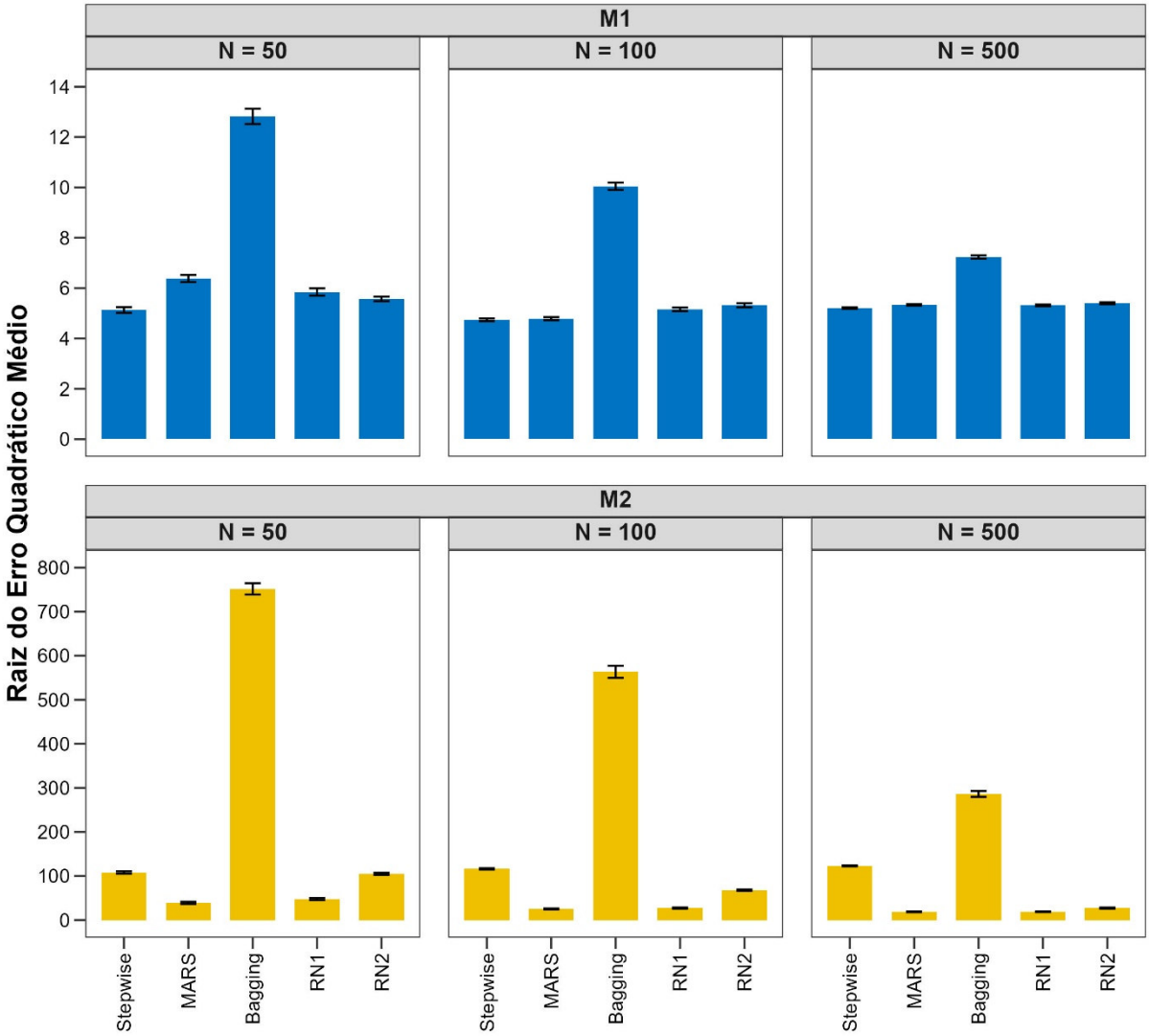


Fonte: Autores (2025).

A metodologia *Bagging*, baseada em árvores de decisão, se destacou como sendo a metodologia que apresentou menor estabilidade em amostras menores e valores de REQM mais elevados (Figura 4). Segundo a literatura, apesar de sua robustez e capacidade de capturar interações de forma implícita, os modelos baseados em árvores não permitem a decomposição interpretável dos efeitos de interação (STROBL *et al*., 2024; BREIMAN, 2024).

Os métodos *Stepwise* e *MARS* se destacam por realizarem, automaticamente, a seleção de variáveis, apresentando modelos reduzidos. No entanto, diferentemente do *MARS*, o *Stepwise* tende a ser mais limitado em contextos com interações, uma vez que não as inclui automaticamente no processo de modelagem (JAMES *et al*., 2021).

FIGURA 4 - Estimativa da raiz dos erros quadráticos médios para as metodologias *Stepwise*, *MARS*, *Bagging* e Redes Neurais (RN1 e RN2), juntamente com os respectivos erros-padrão da média (EPM), considerando os cenários M1 e M2 em 10 repetições.



Fonte: Autores (2025).

Nos dois cenários analisados, observou-se uma redução dos valores de REQM com o aumento do tamanho amostral (Figura 4), o que está de acordo com estudos que associam maiores amostras à melhoria na capacidade preditiva de modelos estatísticos e de aprendizado de máquina (CHENG, 2024).

Dentre as metodologias avaliadas, *MARS* e redes neurais, destacaram-se por apresentarem os menores valores de REQM, indicando maior precisão preditiva. Estes resultados são coerentes com a literatura, que evidencia a capacidade dessas abordagens em modelar relações não lineares e capturar interações entre variáveis, como no cenário M2, caracterizado pela presença de efeito de interação entre as variáveis X_1 e X_2 (FRIEDMAN, 1991; ZHOU, 2021).

Por fim, a semelhança entre os resultados obtidos pelas diferentes metodologias pode estar associada à baixa complexidade estrutural dos modelos

utilizados na simulação dos dados, caracterizados por relações predominantemente lineares e com pouca interação entre variáveis. Em tais contextos, diferentes técnicas estatísticas e de aprendizado de máquina tendem a convergir na identificação das variáveis mais importantes, conforme sugerido por estudos que mostram menor variação no desempenho dos métodos em cenários menos complexos (CHENG, 2024).

CONCLUSÕES

Os resultados indicam que os métodos avaliados foram, em geral, eficazes na identificação das variáveis mais importantes. Observou-se que diferentes metodologias podem apresentar desempenhos similares em cenários de baixa complexidade estrutural, com relações predominantemente lineares e poucas interações. A escolha metodológica deve considerar os objetivos da pesquisa: quando há interesse na interpretação de interações, o MARS se destaca por explicitar esses efeitos; já em contextos em que a capacidade preditiva é prioritária, MARS e RNA's avaliadas pelos métodos de Garson e Olden apresentaram maior CP, com menores valores de REQM. Os resultados também reforçam que maiores tamanhos amostrais aumentam a estabilidade das estimativas e a confiabilidade das predições. Assim, recomenda-se alinhar a metodologia ao tipo de informação desejada, preditiva ou interpretativa, considerando a complexidade dos dados e o tamanho da amostra disponível.

REFERÊNCIAS

AKAIKE, H. **Information theory and an extension of the maximum likelihood principle**. In: PETROV, B. N.; CSAKI, F. (Ed.). Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 1973. p. 267–281. DOI: https://doi.org/10.1007/978-1-4612-1694-0_15.

AL-SUDANI, Z. A.; SALIH, S. Q.; SHARAFATI, A.; YASEEN, Z. M. Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation. **Journal of Hydrology**, v. 573, p. 1–12, 2019. DOI: https://doi.org/10.1016/j.jhydrol.2019.03.004.

BERNARDI, R. E.; GÜNTHER, F. RSNNS: Neural Networks using the Stuttgart Neural Network Simulator (SNNS). **R package** version 0.4-17, 2022. Disponível em: https://cran.r-project.org/package=RSNNS. Acesso em: 09 maio 2025.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, p. 123–140, 1996. DOI: https://doi.org/10.1007/BF00058655.

BREIMAN, L.; CUTLER, A.; LIAW, A.; WIENER, M. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. **R package** Versão 4.7-1.2. 2024. DOI: 10.32614/CRAN.package.randomForest.

CELERI, M. D.; COSTA, W.G.; NASCIMENTO, A.C.C.; AZEVEDO, C.F.; CRUZ, C.D. et al.; Multivariate Adaptive Regression Splines Enhance Genomic Prediction of Non-Additive Traits. **Agronomy**, v. 14, n. 10, p. 2234, 2024. DOI: https://doi.org/10.3390/agronomy14102234.

CHENG, X. A comprehensive study of feature selection techniques in machine learning models. **Insights in Computer Signals and Systems**, v. 1, n. 1, 2024. DOI: http://dx.doi.org/10.2139/ssrn.5154947.

CRUZ, C. D.; NASCIMENTO, M. **Inteligência Computacional Aplicada ao Melhoramento Genético**. Viçosa: Editora UFV, 2018. 414 p. ISBN: 9788572696067. Disponível em: https://www.editoraufv.com.br/produto/inteligencia-computacional-aplicada-ao-melhoramento-genetico/1114585.

FRIEDMAN, J. H. Multivariate adaptive regression splines. **The Annals of Statistics**, v. 19, n. 1, p. 1–67, 1991. DOI: https://doi.org/10.1214/aos/1176347963.

GARSON, G. D. Interpreting neural network connection weights. **Artificial Intelligence Expert**, v. 6, p. 46–51, 1991. DOI: https://dl.acm.org/doi/abs/10.5555/129449.129452.

GEVREY, M.; DIMOPOULOS, I.; LEK, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. **Ecological Modelling**, v. 160, n. 3, p. 249–264, 2003. DOI: https://doi.org/10.1016/S0304-3800(02)00257-0.

GOH, A. T. C. Back-propagation neural networks for modeling complex systems. **Artificial Intelligence in Engineering**, v. 9, n. 3, p. 143–151, 1995. DOI: https://doi.org/10.1016/0954-1810(94)00011-S.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. New York: Springer, 2009. DOI: https://doi.org/10.1007/978-0-387-84858-7.

HUANG, H.; JI, X.; XIA, F.; HUANG, S.; SHANG, X.; CHEN, H.; ZHANG, M.; DAHLGREN, R.A.; MEI, K. Multivariate Adaptive Regression Splines for Estimating Riverine Constituent Concentrations. **Hydrological Processes**. *34*, 1213–1227, 2020. DOI: https://doi.org/10.1002/hyp.13669

IBRAHIM, E. A.; SALIFU, D.; MWALILI, S.; DUBOIS, T.; COLLINS, R.; TONNANG, H. E. Z. An expert system for insect pest population dynamics prediction. **Computers and Electronics in Agriculture,** v. 198, p. 107124, 2022. DOI: https://doi.org/10.1016/j.compag.2022.107124.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning: with applications in R**. 2. ed. New York: Springer, 2021. DOI: https://doi.org/10.1007/978-1-0716-1418-1.

LUNDBERG, S. M.; ERION, G.; CHEN, H.; DeGRAVE, A.; PRUTKIN, J.M.; et al. From local explanations to global understanding with explainable AI for trees. **Nature Machine Intelligence**, v. 2, p. 56–67, 2020. DOI: https://doi.org/10.1038/s42256-019-0138-9.

MILBORROW, S. Earth: Multivariate Adaptive Regression Splines. Versão 5.4.0. 2025. Disponível em: https://cran.r-project.org/package=earth. Acesso em: 12 maio 2025.

MOLNAR, C. **Interpretable machine learning: a guide for making black box models explainable**. 2.ed. Christoph Molnar, 2022. Disponível em: https://christophmolnar.com/books/interpretable-machine-learning. Acesso em: 25 maio 2025.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. 6. ed. New York: Wiley, 2021. 822 p. Disponível em: https://www.wiley.com/en-us/Introduction+to+Linear+Regression+Analysis%2C+6th+Edition-p-9781119578727. Acesso em: 05 maio 2025.

NASCIMENTO, M.; AZEVEDO, C. F.; NASCIMENTO, A. C. C.; CRUZ, C. D. **Abordagens biométricas para reconhecimento de padrões, classificação e predição nas ciências agrárias**. Londrina: Mecenas, 2024. 430 p. ISBN: 9786555860744. Disponível em: https://editoramecenas.com.br/livro/abordagens-biometricas. Acesso em: 05 maio 2025.

OLDEN, J. D.; JACKSON, D. A. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. **Ecological Modelling**, v. 154, p. 135–150, 2002. DOI: https://doi.org/10.1016/S0304-3800(02)00064-9.

OLDEN, J. D.; JOY, M. K.; DEATH, R. G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. **Ecological Modelling**, v. 178, p. 389–397, 2004. DOI: https://doi.org/10.1016/j.ecolmodel.2004.03.013.

R CORE TEAM. R: a language and environment for statistical computing. Vienna: **R Foundation for Statistical Computing**, 2023. Disponível em: https://www.R-project.org/. Acesso em: 10 maio 2025.

RIPLEY, B.; VENABLES, W. N. MASS: Support Functions and Datasets for Venables and **Ripley's MASS**. Versão 7.3-65. 2025. Disponível em: https://cran.r-project.org/package=MASS. Acesso em: 12 maio 2025.

SILVA JÚNIOR, A.C.; SANT’ANNA, I.C.; SILVA, G.N.; CRUZ, C.D.; NASCIMENTO, M.; et al.; Computational intelligence and machine learning to study the importance of characteristics in flood-irrigated rice. **Acta Scientiarum**. Agronomy, v. 45, e57209, 2022. DOI: https://doi.org/10.4025/actasciagron.v45i1.57209.

STROBL, C., ROTHACHER, Y., THEILER, S. & HENNINGER, M. Detecting interactions with random forests: a comment on Gries’ words of caution and suggestions for improvement. **Corpus Linguistics and Linguistic Theory,** 2024. DOI: https://doi.org/10.1186/1471-2105-9-307.

VENABLES, W. N.; RIPLEY, B. D. *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. **R package** version 7.3-19, 2023. Disponível em: https://cran.r-project.org/package=nnet. Acesso em: 09 maio 2025.

ZHOU, Z.; **Machine learning**. Tradução de Shaowu Liu. 1. ed. Singapura: Springer, 2021. DOI: https://doi.org/10.1007/978-981-15-1967-3.