Classificação binária de cogumelos por meio de árvore CHAID: uma aplicação ao Secondary Mushroom Dataset

Leonardo de Oliveira Maltez[1], Marcello Montillo Provenza[1,2], Igor C. A. Lima[1]

[1]*Universidade do Estado do Rio de Janeiro, Rio de Janeiro, RJ, Brasil (igor.campos@ime.uerj.br)*
[2] *Universidade do Estado do Rio de Janeiro, São Gonçalo, RJ, Brasil*

Resumo: Este trabalho avaliou uma árvore CHAID para classificação binária de cogumelos no Secondary Mushroom Dataset, considerando a classe tóxica como evento de interesse. Após pré-processamento e validação por conjunto de teste, o modelo apresentou desempenho preditivo elevado, com sensibilidade de 0,9999, especificidade de 0,9985, AUC de 0,9993 e apenas um falso negativo no conjunto de teste, indicando potencial como ferramenta auxiliar de triagem, embora recomende-se validação externa.

Palavras-chave: CHAID; cogumelos venenosos; classificação binária; matriz de confusão; aprendizado de máquina.

INTRODUÇÃO

Cogumelos são corpos frutíferos macroscópicos de certos fungos, isto é, estruturas visíveis associadas à reprodução e à dispersão de esporos, e o uso popular do termo costuma privilegiar as formas comestíveis, enquanto formas não comestíveis ou venenosas são frequentemente tratadas como um subconjunto de maior risco alimentar (Britannica, 2026). Como pertencem ao reino Fungi, os cogumelos não devem ser descritos botanicamente como plantas, mas como organismos fúngicos, grupo que inclui leveduras, bolores, ferrugens, mofos e formas macroscópicas como os próprios cogumelos (*ibidem*).

Sua importância alimentar, econômica e sociocultural é ampla, pois fungos comestíveis silvestres são coletados para alimentação, comércio e geração de renda em diversos países, havendo registros de grande diversidade de espécies utilizadas por comunidades humanas, inclusive em contextos de subsistência e de mercados locais ou internacionais (Boa, 2004). Essa relevância, contudo, convive com perigo real, porque espécies comestíveis e espécies tóxicas podem ser confundidas por semelhança morfológica, e intoxicações por cogumelos podem variar de sintomas gastrointestinais a síndromes graves, com insuficiência hepática e morte em casos associados a amatoxinas (Garcia et al., 2015; Li, 2019). Entre os exemplos mais conhecidos está *Amanita phalloides*, espécie frequentemente associada a envenenamentos fatais por amatoxinas, o que torna a correta discriminação entre cogumelos comestíveis e venenosos uma questão de segurança alimentar e saúde pública, além de um problema taxonômico e analítico (Garcia et al., 2015).

A classificação de cogumelos e a discriminação entre formas comestíveis, venenosas ou potencialmente tóxicas têm sido abordadas por diferentes estratégias, incluindo aprendizagem de máquina baseada em atributos morfológicos, visão computacional, espectroscopia, cromatografia e métodos de detecção de toxinas. No campo de bases de dados e classificação supervisionada, Wagner, Heider e Hattab (2021) construíram e curaram dados destinados à classificação binária de cogumelos em comestíveis e venenosos, avaliando classificadores como Naive Bayes, regressão logística, análise discriminante linear e Random Forest. Em visão computacional, Ketwongsa, Boonlue e Kokaew (2022) propuseram um modelo de aprendizagem profunda baseado em uma versão modificada da AlexNet para classificar imagens de cogumelos em venenosos e comestíveis, comparando a proposta com arquiteturas convolucionais conhecidas. Em espectroscopia, Chen et al. (2023) revisaram aplicações de UV-Vis e infravermelho combinadas com quimiometria para discriminação e avaliação de cogumelos bolete silvestres comestíveis, incluindo identificação de espécies, origem, tempo de armazenamento, detecção de fraude e avaliação de propriedades antioxidantes. Também por via espectroscópica, Chen et al. (2023) aplicaram FT-NIR combinada com DD-SIMCA e Random Forest para discriminar comestibilidade e espécies de boletes silvestres, relatando o potencial da abordagem para triagem de espécies potencialmente venenosas e para discriminação taxonômica. Em química analítica, Carvalho et al. (2014) aplicaram GC-MS a 22 espécies de cogumelos silvestres, incluindo espécies comestíveis, tóxicas e potencialmente tóxicas, usando perfis químicos para

caracterização e diferenciação das amostras. Por fim, Barbosa et al. (2022) desenvolveram métodos por HPLC-UV-EC e HPLC-DAD-MS para determinação de α- e β-amanitina em cogumelos silvestres, enquanto Lu et al. (2024) propuseram método LC-MS/MS para detecção simultânea de múltiplas toxinas de cogumelos, mostrando que a discriminação entre cogumelos seguros e perigosos também pode ser apoiada pela identificação direta de compostos venenosos.

O detector automático de interação baseado em qui-quadrado (*CHi-square Automatic Interaction Detection*, CHAID), proposto por Kass (1980) é uma técnica de árvore de decisão voltada sobretudo para problemas em que a variável resposta é categórica. O objetivo é particionar a amostra em subgrupos mutuamente exclusivos e exaustivos que diferem quanto à distribuição da variável resposta. É derivada do AID (*Automatic Interaction Detection*), proposto por Morgan e Sonquist (1963).

O CHAID serve para segmentação, classificação, exploração de interações e construção de regras interpretáveis. Sua lógica é recursiva: em cada nó da árvore, o algoritmo testa quais variáveis explicativas apresentam associação estatisticamente significativa com a variável resposta e seleciona aquela que produz a melhor partição. Diferentemente de muitas árvores binárias, o CHAID pode produzir divisões múltiplas, isto é, um nó pode se abrir em mais de dois ramos. Kass enfatiza justamente essas diferenças em relação ao AID: uso embutido de testes de significância, seleção do preditor mais significativo e divisões múltiplas.

Para cada preditor categórico, o algoritmo compara categorias que não diferem significativamente quanto à distribuição da resposta. Categorias estatisticamente semelhantes podem ser fundidas. Depois das fusões admissíveis, calcula-se uma significância ajustada, frequentemente com correção de Bonferroni ou variantes, para controlar comparações múltiplas. Em seguida, entre os preditores candidatos, escolhe-se aquele com associação mais significativa com a resposta. Se o valor-p ajustado ultrapassar o limiar de divisão definido pelo analista, o nó não é dividido. Critérios operacionais usuais incluem nível de significância para divisão, nível de significância para mesclagem, número máximo de iterações, tamanho mínimo do nó pai, tamanho mínimo dos nós filhos e restrições sobre frequências esperadas.

Até onde foi possível verificar nas bases e mecanismos de busca consultados, não foram localizados estudos aplicando especificamente o algoritmo CHAID ao *Secondary Mushroom Dataset* (Wagner, Heider e Hattab, 2021). Assim, o presente trabalho explora essa lacuna, aplicando uma árvore CHAID a uma base pública, recente e apropriada para classificação binária com preditores majoritariamente categóricos. Assim, este trabalho tem por objetivo testar o algoritmo CHAID nesta base de dados, bem como avaliar seu desempenho preditivo.

MATERIAL E MÉTODOS

Na utilização de classificadores, diferentes métricas são utilizadas para validar seu desempenho. Essas métricas, denominadas como parâmetros de mérito, são construídas a partir da matriz de confusão (tab. 1), que é uma tabela que compara a classificação real da classe de interesse com a classificação predita pelo modelo. No presente trabalho, considerando positivo como cogumelo venenoso, ela permite contar quantos casos foram corretamente ou incorretamente classificados.

Tabela 1. Matriz de confusão.

Valor real	Valor predito	
	Positivo	Negativo
Positivo	VP	FN
Negativo	FP	VN

Fonte: Adaptado de Kuhn e Johnson, 2016.

onde VP são os verdadeiros positivos (cogumelos venenosos corretamente classificados como venenosos); VN são os verdadeiros negativos (cogumelos não venenosos corretamente classificados como não venenosos); FP são os falsos positivos (cogumelos não venenosos classificados indevidamente como venenosos); e FN são os falsos negativos (cogumelos venenosos classificados incorretamente como não venenosos). No problema de identificação de cogumelos venenosos, o erro mais crítico é o falso negativo, pois representa liberar como não venenoso um cogumelo que, na realidade, é venenoso.

A prevalência (eq. 1) é a proporção amostras de interesse (no caso, de cogumelos realmente venenosos) na amostra:

$$\text{Prev} = \frac{\text{VP} + \text{FN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}} \qquad (1)$$

A sensibilidade (eq. 2) mede a capacidade do modelo de detectar os venenosos (Kuhn e Johnson, 2016, p. 256):

$$\text{Sens} = \text{TPR} = \frac{\text{VP}}{\text{VP} + \text{FN}} \qquad (2)$$

A especificidade (eq. 3) mede a capacidade de reconhecer corretamente os não venenosos (*ibidem*):

$$\text{Esp} = \frac{\text{VN}}{\text{VN} + \text{FP}} \quad (3)$$

A precisão (eq. 4) ou valor preditivo positivo empírico mede, entre os classificados como venenosos, quantos eram realmente venenosos:

$$\text{Prec} = \frac{\text{VP}}{\text{VP} + \text{FP}} \quad (4)$$

A exatidão (eq. 5) ou acurácia mede a proporção total de classificações corretas:

$$\text{Exat} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \quad (5)$$

O coeficiente de correlação de Matthews (eq. 6) é uma medida global de associação entre classe real e classe predita, levando em conta simultaneamente VP, VN, FP e FN, sendo especialmente útil quando há desequilíbrio entre classes (Matthews, 1975):

$$\text{MCC} = \frac{\text{VP} \cdot \text{VN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{VP} + \text{FP})(\text{VP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (6)$$

O índice Kappa de Cohen (eq. 7) mede a concordância entre classificação observada, p_o, e predita corrigindo a concordância esperada ao acaso, p_e (Cohen, 1960 *apud* Kuhn e Johnson, 2016, p. 255):

$$\kappa = \frac{p_o + p_e}{1 - p_e} \quad (7)$$

A equação 8, proposta por Van Rijsbergen (1979), na avaliação de sistemas de recuperação de informação, permite a origem de duas métricas: o escore F_1 e o escore F_2.

$$F_\beta = \frac{(1 + \beta^2)\text{Precisão} \cdot \text{Sensibilidade}}{\beta^2 \cdot \text{Precisão} + \text{Sensibilidade}} \quad (8)$$

O escore F_1 (eq. 9) é a média harmônica entre precisão e sensibilidade, equilibrando falsos positivos e falsos negativos (Van Rijsbergen, 1979):

$$F_1 = \frac{2 \cdot \text{Precisão} \cdot \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \quad (9)$$

O escore F_2 (eq. 10) também combina precisão e sensibilidade, mas atribui maior peso à sensibilidade; por isso, é muito apropriado quando é mais grave deixar de detectar um cogumelo venenoso:

$$F_2 = \frac{5 \cdot \text{Precisão} \cdot \text{Sensibilidade}}{4 \cdot \text{Precisão} + \text{Sensibilidade}} \quad (10)$$

O índice de Fowlkes-Mallows (eq. 11) é a média geométrica entre precisão e sensibilidade (Fowlkes-Mallows, 1983):

$$\text{FMI} = \sqrt{\text{Precisão} \times \text{Sensibilidade}} \quad (11)$$

A AUC (eq. 12), é a área sob a curva ROC e mede a capacidade global do modelo de separar cogumelos venenosos e não venenosos ao longo de diferentes pontos de corte (Bamber, 1975):

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\,d(\text{FPR}) \quad (12)$$

O PPV teórico (eq. 13) é a probabilidade de um cogumelo ser realmente venenoso dado que o modelo o classificou como venenoso, calculada a partir de sensibilidade, especificidade e prevalência (Kuhn e Johnson, 2016, p. 258):

$$\text{VPP} = \frac{\text{Sens} \times \text{Prev}}{(\text{Sens} \times \text{Prev}) + (1 - \text{Esp})(1 - \text{Prev})} \quad (13)$$

O VPN teórico (eq. 14) é a probabilidade de um cogumelo ser realmente não venenoso dado que o modelo o classificou como não venenoso, também dependendo de sensibilidade, especificidade e prevalência (Kuhn e Johnson, 2016, p. 258):

$$\text{VPN} = \frac{\text{Esp} \times (1 - \text{Prev})}{(1 - \text{Sens}) \times \text{Prev} + \text{Esp}(1 - \text{Prev})} \quad (14)$$

O índice J de Youden (eq. 15) mede a capacidade global do classificador de separar corretamente os cogumelos venenosos dos não venenosos, combinando sensibilidade e especificidade em uma única medida (Youden, 1950 *apud* Kuhn e Johnson, 2016, p. 257):

$$J = \text{Sens} + \text{Esp} - 1 \quad (15)$$

Se J=1, o classificador é perfeito; se J=0, ele não discrimina melhor do que uma decisão sem utilidade prática; e valores próximos de 1 indicam excelente capacidade discriminatória. No contexto de cogumelos venenosos, um J elevado significa que o modelo combina alta capacidade de detectar os venenosos com alta capacidade de reconhecer corretamente os não venenosos.

A base de dados foi extraída do repositório de dados UC Irvine Machine Learning Repository (UCI, 2026) e foi publicada originalmente por Wagner, Heider e Hattab (2021). A base é constituída por 61.069 instâncias e 21 variáveis (entre variáveis contínuas e categóricas). A variável-alvo é constituída por duas classes, comestível (27.181 instâncias) e venenoso (33.888 instâncias), na proporção de 44,51% e 55,49%, respectivamente.

O pré-processamento incluiu conversão das variáveis métricas para formato numérico, conversão das variáveis nominais para fator e substituição de categorias ausentes, vazias ou codificadas com o símbolo de interrogação (?) pela categoria Unknown. Foram excluídas variáveis explicativas com mais de 50% de valores ausentes, preservando-se a variável resposta. Para as variáveis métricas remanescentes, valores ausentes foram imputados pela mediana da própria variável; para as variáveis nominais, valores ausentes ou vazios foram imputados pela moda simples observada. Registros com classe ausente foram removidos antes da modelagem.

A amostra foi particionada de forma estratificada em conjunto de treinamento e conjunto de teste, com semente aleatória igual a 123. As amostras foram divididas em 80% das observações para treinamento e 20% para teste, por meio da função createDataPartition do pacote caret. Após a divisão, os níveis das variáveis nominais no conjunto de teste foram alinhados aos níveis observados no conjunto de treinamento; categorias ausentes ou não observadas no treinamento foram substituídas por Unknown quando esse nível estava disponível, ou pelo primeiro nível do fator de treinamento caso contrário.

Para o ajuste da árvore CHAID, não foi utilizada codificação dummy. As variáveis explicativas foram preparadas diretamente como fatores, conforme exigido pelo algoritmo implementado. As variáveis métricas foram discretizadas em cinco faixas ordinais, definidas por quantis, usando a função cut com include.lowest = TRUE. Quando a variável não apresentava número suficiente de pontos de corte distintos, seus valores foram convertidos em categorias fatoriais, com tratamento explícito de ausências pela categoria Unknown. Ao final do pré-processamento específico do CHAID, todas as variáveis explicativas foram verificadas quanto à classe factor, interrompendo-se a rotina caso alguma variável ainda não estivesse nesse formato.

O desempenho do modelo CHAID foi avaliado por validação cruzada k-fold no conjunto de treinamento, com k = 5. A rotina utilizou os folds criados pela função createFolds, com returnTrain = TRUE, e a função trainControl do pacote caret. Para o CHAID, foram salvas as predições finais da validação cruzada, não foram usadas probabilidades de classe na função-resumo da validação cruzada, e a otimização foi conduzida com base no escore F_2. A escolha do F_2 se justifica porque essa medida atribui maior peso à sensibilidade do que à precisão, o que é coerente com o problema aplicado, no qual deixar de identificar cogumelos venenosos representa erro mais crítico do que classificar indevidamente um cogumelo comestível como venenoso.

A árvore CHAID foi ajustada no caret com method = "chaid". A grade de hiperparâmetros avaliou alpha2 nos valores 0,01, 0,05 e 0,10; alpha3 fixado em -1; e alpha4 nos valores 0,01, 0,05 e 0,10. O melhor modelo foi definido pela maximização do F_2 na validação cruzada. Para reduzir o tempo de processamento, a rotina registrou computação paralela para a etapa do CHAID por meio de cluster PSOCK, com número de núcleos limitado a no máximo quatro.

Após o ajuste, o modelo selecionado foi aplicado ao conjunto de teste independente. As predições foram comparadas aos valores observados por meio de matriz de confusão, assumindo poisonous como evento positivo.

Foram obtidas a prevalência, a exatidão, a sensibilidade, a especificidade, a precisão, os escores F_1 e F_2, o índice J de Youden, o coeficiente de correlação de Matthews (MCC), o coeficiente Kappa de Cohen, bem como os valores preditivos positivo e negativo teóricos em função da prevalência observada no conjunto de teste, da sensibilidade e da especificidade. A AUC foi calculada somente quando havia probabilidades preditas válidas para a classe poisonous.

A análise de dados foi feita em ambiente RStudio (Posit, 2025), com os pacotes tidyverse (Wickham *et al.*, 2019), caret (Kuhn, 2008), pROC (Robin *et al.*, 2011), naivebayes (Majka, 2024), doParallel (Microsoft e Weston, 2022), foreach (Microsoft e Weston, 2022), ggplot2 (Wickham, 2016), lattice (Sarkar, 2008) e CHAID (The FoRt Student Project Team, 2015).

Foi empregado um notebook Lenovo com processador 12th Gen Intel(R) Core(TM) i7-12700H (2.30 GHz); com 32,0 GB de memória RAM e placa gráfica NVIDIA GeForce RTX 3060 Laptop GPU (6 GB).

RESULTADOS E DISCUSSÃO

Após a preparação da base de dados, foram atribuídas 48.856 amostras para o conjunto de treinamento e 12.213 amostras para o conjunto de teste.

A tabela 2 mostra a matriz de confusão, para a etapa de validação do modelo.

Tabela 2. Matriz de confusão: etapa de validação.

Valor real	Valor predito	
	Positivo	Negativo
Positivo	VP = 6776	FN = 1
Negativo	FP = 8	VN = 5428

Fonte: Os autores, 2026.

A matriz de confusão indica um desempenho elevado do modelo na etapa de validação. O modelo classificou corretamente 6776 casos positivos (cogumelos venenosos) e 5428 casos negativos, totalizando 12204 acertos em 12213 observações. Houve apenas 9 erros de classificação, sendo 1 falso negativo e 8 falsos positivos.

A partir da tabela 2, foram calculados os parâmetros de mérito, mostrados na tabela 3.

Tabela 3. Parâmetros de mérito, etapa validação.

Parâmetro	Resultado
Prevalência	0,554900516
Sensibilidade	0,999852442
Especificidade	0,99852833
Precisão	0,998820755
Exatidão	0,99926308
J de Youden	0,998380772
Escore F_1	0,999336332
Escore F_2	0,999645934
FMI	0,999336465
AUC	0,999333519
MCC	0,998508657
Kappa	0,998507985
PPV_teórico	0,998820755
NPV_teórico	0,999815804

Fonte: Os autores, 2026.

Os resultados indicam desempenho preditivo elevado na etapa de validação. A matriz de confusão mostra 6.776 verdadeiros positivos, isto é, cogumelos venenosos corretamente identificados como venenosos, e 5.428 verdadeiros negativos, correspondentes aos cogumelos não venenosos corretamente classificados como não venenosos. Foram observados apenas 8 falsos positivos e 1 falso negativo em um total de 12.213 observações, o que representa uma taxa global de erro baixa. No contexto de segurança alimentar, o resultado mais relevante é justamente o falso negativo, pois esse erro corresponde a classificar como não venenoso um cogumelo que, na realidade, é venenoso. O fato de ter ocorrido apenas um falso negativo indica que o classificador praticamente não deixou passar cogumelos perigosos.

A sensibilidade de 0,999852442 mostra que o modelo identificou corretamente aproximadamente 99,985% dos cogumelos venenosos. Esse resultado é particularmente importante porque, nesse tipo de aplicação, a prioridade deve ser minimizar a probabilidade de um cogumelo venenoso ser classificado como seguro. A especificidade de 0,99852833 também foi elevada, indicando que aproximadamente 99,853% dos cogumelos não venenosos foram corretamente reconhecidos como não venenosos. Os 8 falsos positivos significam que alguns cogumelos seguros foram classificados como venenosos, erro que pode gerar descarte indevido, mas que é menos grave do que liberar um cogumelo venenoso como comestível.

A precisão de 0,998820755 e o valor preditivo positivo teórico de mesmo valor indicam que, entre os cogumelos classificados como venenosos pelo modelo, cerca de 99,882% eram realmente venenosos.

Já o valor preditivo negativo de 0,999815804 indica que, entre os cogumelos classificados como não venenosos, aproximadamente 99,982% estavam corretamente classificados. Esses dois resultados mostram que tanto a predição positiva quanto a predição negativa foram altamente confiáveis, embora a interpretação do valor preditivo negativo deva ser feita com atenção, pois o único falso negativo observado está justamente nesse grupo de amostras classificadas como não tóxicas.

A acurácia de 0,99926308 indica que aproximadamente 99,926% das classificações totais foram corretas. Entretanto, nesse problema, a acurácia isoladamente não deve ser a principal referência, pois a gravidade dos erros é assimétrica: errar um cogumelo venenoso como não venenoso é mais crítico do que errar um cogumelo não venenoso como venenoso. Por isso, métricas como sensibilidade, escore F_2, valor preditivo negativo e índice de Youden são mais informativas para avaliar a utilidade prática do modelo.

O escore F_1 de 0,999336332 confirma o equilíbrio entre precisão e sensibilidade, enquanto o escore F_2 de 0,999645934 é especialmente relevante, pois atribui maior peso à sensibilidade. Esse resultado reforça que o modelo teve ótimo desempenho justamente no critério mais importante para a aplicação (a identificação dos cogumelos venenosos).

O índice de Youden de 0,998380772 evidencia excelente capacidade de discriminação entre cogumelos venenosos e não venenosos, pois valores próximos de 1 indicam combinação quase ideal entre sensibilidade e especificidade.

A figura 1 mostra a curva ROC, na etapa de validação, para o método proposto.

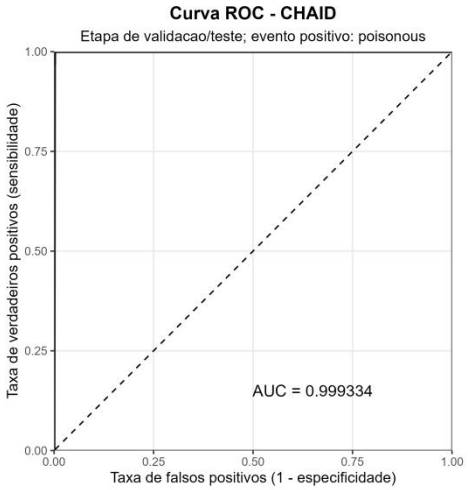


Figura 1. Curva ROC do método CHAID, na etapa de validação.

Observa-se, na figura 1, desempenho discriminatório elevado, com AUC igual a 0,999334. Esse valor, muito próximo de 1, indica que o modelo possui elevada capacidade de separação entre cogumelos venenosos e comestíveis, ao longo de diferentes limiares de decisão. A curva ROC situa-se acima da linha diagonal de referência, que representa um classificador aleatório, evidenciando que o CHAID apresenta alta sensibilidade associada a baixa taxa de falsos positivos. Esse resultado é compatível com a matriz de confusão da etapa de validação, na qual o modelo produziu número reduzido de erros de classificação.

De modo complementar, o MCC de 0,998508657 e o Kappa de 0,998507985 indicam concordância quase perfeita entre as classificações observadas e preditas, levando em conta simultaneamente acertos e erros nas duas classes.

O índice de Fowlken-Mallows (FMI) de 0,999336465, também confirma o desempenho praticamente máximo do classificador na identificação da classe positiva.

Por fim, destaca-se o tempo de processamento foi de aproximadamente 30 minutos. Esse tempo é relativamente elevado, mas aceitável considerando a etapa de validação e ajuste deste modelo. Entretanto, o tempo deve ser considerado na avaliação de custo computacional do procedimento, especialmente em cenários com reamostragem, ajuste de hiperparâmetros e bases de maior dimensão.

CONCLUSÃO

Os resultados evidenciam que o classificador apresentou desempenho elevado na identificação de cogumelos venenosos, combinando alta sensibilidade, elevada especificidade, baixa frequência de classificações incorretas e forte concordância entre as classes observadas e preditas. Considerando a classe tóxica como evento de interesse, destaca-se a capacidade do modelo em reconhecer praticamente todos os casos positivos, aspecto fundamental em problemas de segurança alimentar, nos quais o falso negativo representa o erro de maior gravidade prática. Ainda assim, a ocorrência de um falso negativo indica que o desempenho, embora próximo do ideal na base avaliada, não deve ser interpretado como garantia absoluta de segurança.

Dessa forma, os resultados sustentam o potencial do modelo como ferramenta auxiliar para triagem e classificação de cogumelos, mas recomenda-se validação externa, preferencialmente com bases independentes e condições reais de aplicação, antes de seu uso como instrumento decisório definitivo.

AGRADECIMENTOS

Leonardo de Oliveira Maltez e Marcello Montillo Provenza agradecem a bolsa do Programa de Inovação em Docência da Universidade do Estado do Rio de Janeiro (PID/UERJ), Protocolo PID25-nW5IJq.

REFERÊNCIAS

BAMBER, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology, v. 12, n.º 4, p. 387-415, 1975.

BOA, E. Wild edible fungi: a global overview of their use and importance to people. Rome: FAO, 2004. Non-Wood Forest Products 17.

BARBOSA, I. *et al.* Amanitins in Wild Mushrooms: The Development of HPLC-UV-EC and HPLC-DAD-MS Methods for Food Safety Purposes. Foods, v. 11, n.º 23, 3929, 2022.

BRITANNICA, The Editors of Encyclopaedia. Mushroom. Encyclopaedia Britannica, 2026.

CARVALHO, L. M. *et al.* Non-targeted and targeted analysis of wild toxic and edible mushrooms using gas chromatography-ion trap mass spectrometry. Talanta, v. 118, p. 292-303, 2014.

CHEN, J. *et al*. Application of UV-Vis and Infrared Spectroscopy on Wild edible bolete mushrooms discrimination and evaluation: a review. Critical Reviews in Analytical Chemistry, v. 53, n. 4, p. 852-868, 2023.

CHEN, J. et al. Edibility and species discrimination of wild bolete mushrooms using FT-NIR spectroscopy combined with DD-SIMCA and RF models. LWT, v. 182, 114701, 2023.

COHEN, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, v. 20, n. 1, p. 37-46, 1960.

FOWLKES, E. B.; MALLOWS, C. L.. A method for comparing two hierarchical clusterings. Journal of the American Statistical Association, vol. 78, n.º 383, p. 553-569, 1983.

GARCIA, J. et al. Amanita phalloides poisoning: Mechanisms of toxicity and treatment. Food and Chemical Toxicology, v. 86, p. 41-55, 2015.

KASS G. V.. An exploratory technique for investigating large quantities of categorical data. Applied Statistics, vol. 29, n.º 2, p. 119-127, 1980.

KETWONGSA, W.; BOONLUE, S.; KOKAEW, U. A new deep learning model for the classification of poisonous and edible mushrooms based on improved AlexNet convolutional neural network. Applied Sciences, v. 12, n.º 7, 3409, 2022.

KUHN, M. Building predictive models in R using the caret package. Journal of Statistical Software, v. 28, n. 5, p. 1-26, 2008.

KUHN, M.; JOHNSON, K. Applied predictive modelling. New York: Elsevier, 2016.

LI, R. The world's most poisonous mushroom, Amanita phalloides, is growing in BC. British Columbia Medical Journal, v. 61, n.º 1, p. 20-24, 2019.

LU, J. *et al*. Simultaneous determination of multi-class mushroom toxins in mushroom and biological liquid samples using LC-MS/MS. separations, v. 11, n.º 6, 183, 2024.

MAJKA, M. naivebayes: high performance implementation of the Naive Bayes algorithm in R. R package version 1.0.0, 2024.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure, v. 405, n.º 2, p. 442-451, 1975.

MICROSOFT CORPORATION; WESTON, S. doParallel: foreach parallel adaptor for the parallel package. R package version 1.0.17, 2022a.

MICROSOFT CORPORATION; WESTON, S. foreach: provides foreach looping construct. R package version 1.5.2, 2022b.

MORGAN, J. N.; SONQUIST, J. A. Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Association, v. 58, n. 302, p. 415-434, 1963.

POSIT TEAM. RStudio: integrated development environment for R. Boston: Posit Software, PBC, 2025.

ROBIN, X. *et al*. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, v. 12, artigo 77, 2011.

SARKAR, D. Lattice: multivariate data visualization with R. New York: Springer, 2008.

THE FORT STUDENT PROJECT TEAM; HOTHORN, T. CHAID: CHi-squared automated interaction detection. R package version 0.1-2, R-Forge, 2009.

UC Irvine Machine Learning Repository. Secondary Mushroom. Disponível em: <https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset>. Acesso: 18 maio 2026.

VAN RIJSBERGEN, C. J. Information retrieval, 2nd ed. London: Butterworths, 1979.

WAGNER, D.; HEIDER, D.; HATTAB, G. Mushroom data creation, curation, and simulation to support classification tasks. Scientific Reports, v. 11, 8134, 2021.

WICKHAM, H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag, 2016.

WICKHAM, H. et al. Welcome to the tidyverse. Journal of Open Source Software, v. 4, n. 43, 1686, 2019.

YOUDEN, W. J. Index for rating diagnostic tests. Cancer, v. 3, n.º 1, p. 32-35, 1950.