Classificação Binária de Cogumelos Comestíveis e Venenosos

Victor G. Lima
victor.gl@aluno.ufop.edu.br
Universidade Federal de Ouro Preto
Ouro Preto, Minas Gerais, Brasil

Resumo

A identificação de cogumelos como comestíveis ou venenosos é um importante assunto na micologia. Dada a semelhança morfológica entre as diferentes espécies de cogumelos, a identificação desses seres como comestíveis ou venenosos se torna um desafio. A partir de um dataset contendo informações sobre as características físicas de diversos cogumelos, a classificação binária é capaz de predizer se um cogumelo é comestível ou venenoso. Nesse sentido, o presente trabalho explora técnicas de Mineração de Dados para descrever, pré-processar e encontrar padrões frequentes que possibilitem a identificação desse fungos como comestíveis ou não. Além disso, para lidar com o grande número de instâncias e os diferentes atributos nominais disponíveis, o trabalho utiliza um classificador CatBoost e explora duas modalidades de avaliação experimental: *10-fold cross-validation* e *hold-out*, ambas com buscas de melhor combinação de hiper-parâmetros. Os resultados dos experimentos foram calculados utilizando *Matthews Correlation Coefficient* (MCC), obtendo valores entre 0.98201 e 0.98206.

Keywords

Mushroom, Binary Classification, Pattern Mining

1 Introdução

Cogumelos são estruturas multicelulares que constituem o corpo de frutificação dos fungos, sendo responsáveis pela produção e dispersão dos esporos, que são fundamentais no ciclo de vida desses organismos. A etnomicologia, subdisciplina da etnobotânica, dedica-se ao estudo das interações históricas, culturais e sociais entre os seres humanos e os fungos. Em diversas culturas, os cogumelos comestíveis desempenham um papel significativo na alimentação, sendo amplamente utilizados na culinária tradicional. No entanto, a distinção entre espécies comestíveis e tóxicas é de suma importância, uma vez que a ingestão de cogumelos venenosos pode ocasionar efeitos adversos que variam desde distúrbios gastrointestinais e alterações neuropsíquicas até consequências fatais.

Métodos empíricos são historicamente empregados por diversas sociedades que coletam cogumelos na natureza para fins alimentares e comerciais. No entanto, tais métodos carecem de precisão e podem oferecer riscos à saúde, especialmente devido à semelhança morfológica entre espécies comestíveis e tóxicas. Essa classificação baseia-se majoritariamente em características morfológicas, como as dimensões e formas do chapéu e estipe, além de atributos como coloração, textura e odor. A Figura 1 ilustra as principais estruturas morfológicas utilizadas na identificação de cogumelos.

Com base nessas características, diversos estudos têm explorado a classificação de cogumelos utilizando conjuntos de dados estruturados. Um exemplo é a base *UCI Mushroom*[1], que disponibiliza amostras de cogumelos acompanhadas de suas respectivas descrições morfológicas. Nesse sentido, o presente trabalho tem como objetivo investigar um conjunto de dados contendo um grande número de amostras de cogumelos, aplicando técnicas de mineração de dados para realizar a limpeza e o pré-processamento dos dados, assim como identificar padrões frequentes que possam auxiliar na distinção entre espécies comestíveis e venenosas. A implementação do trabalho e seus resultados estão disponíveis em seu repositório[2].

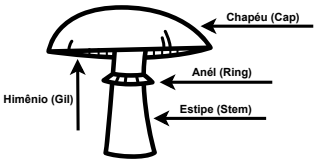


Figura 1: Principais características morfológicas de cogumelos.

2 Trabalhos Relacionados

Para que seja possível o estudo de métodos e técnicas para identificação e classificação de cogumelos, é fundamental que datasets rotulados sejam disponibilizados. Um conjunto que já foi amplamente utilizado foi disponibilizado pela Universidade da California (UCI) em 1987. O dataset apresenta dados de 23 espécies de cogumelos, contendo informações de quais são comestíveis ou não. Entretanto, a base é considerada muito pequena, não refletindo a variedade de características que podem existir nesses seres. Wagner et al. [8] faz disponível uma base secundária, *Secondary Mushroom*[3], contendo 173 espécies de cogumelos presentes em 23 famílias, expandindo o número de amostras e aumentando a variedade de características para classificação. A Tabela 2 apresenta os atributos presentes nas amostras desse dataset e seus possíveis valores.

A partir desses dados, diversos estudos são realizados para classificar cogumelos utilizando técnicas de aprendizado de máquina. Tutuncu et al. [6] utilizam o Secondary Musheoom para comparar diferentes modelos de aprendizado de máquina, incluindo Árvores de Decisão, *Naive Bayes*, *AdaBoost* e *Support Vector Machine* (resultados na Tabela 1. Tutuncu et al. [7] exploram o uso de um classificador XGBoost para a classificação de cogumelos, avaliando com o coeficiente MCC (*Matthews Correlation Coefficient*).

Além da classificação por atributos, existe também a classificação por imagens de cogumelos, que exploram técnicas de Visão Computacional identificar e classificar cogumelos por meio de suas características visuais [3]. Essa abordagem permite que modelos de aprendizado profundo, como redes neurais convolucionais (CNNs),

[1] https://doi.org/10.24432/C5959T

[2] https://github.com/Victorgonl/PoisonousMushrooms/

[3] https://doi.org/10.24432/C5FP5Q

Tabela 1: Resultados apresentados em Tutuncu et al. [6] para o *Secondary Dataset*.

Algoritmo	Accuracy	Recall	Precision	F-Score
Naive Bayes	90.99	98.60	86.04	91.89
Decision Tree	98.82	97.72	100.00	98.85
SVM	99.98	100.00	99.95	99.98
AdaBoost	100.00	100.00	100.00	100.00

extraiam automaticamente padrões visuais relevantes, como textura, forma e cor, facilitando a identificação precisa de espécies mesmo em contextos onde a taxonomia baseada em atributos tabulares seria limitada ou imprecisa. A Figura 2 apresenta um exemplo de dataset composto de imagens de cogumelos para a classificação de espécies comestíveis e venenosas.

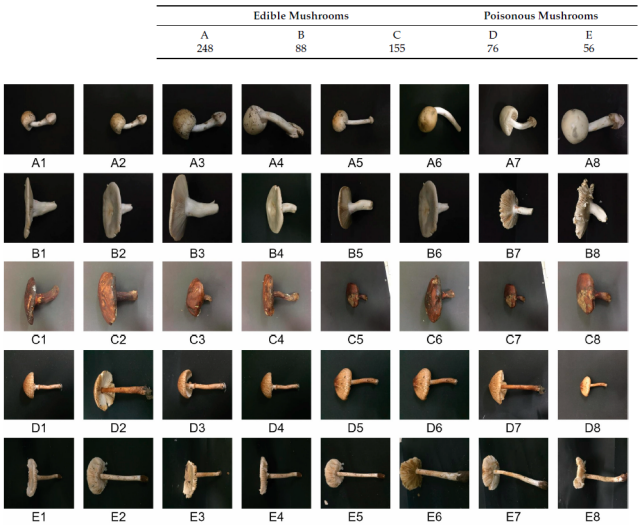


Figura 2: Dataset de imagens de cogumelos disponibilizado por Ketwongsa et al. [3]. A tabela presente na imagem mostra a distribuição de dados no dataset.
As espécies: (A) *Amanita citrina*; (B) *Russula delica*; (C) *Phaeogyroporus portentosus*; (D) *Inocybe rimosa*; (E) *Amanita phalloides*.
Fonte: Ketwongsa et al. [3].

3 Método

Para o presente trabalho, foi utilizado o dataset baseado e expandido no *Secondary Mushroom* e disponibilizado no desafio *Binary Prediction of Poisonous Mushrooms*[4] da plataforma Kaggle. A Tabela 3 apresenta a distribuição dos dados entre as partições de treino e teste. Os atributos do dataset são os mesmos apresentados na Tabela 2. Os valores para a classe objetivo só estão disponíveis para os dados de treino.

O método de análise e processamento do dataset para atingir o objetivo de predição foi dividio em três etapas principais, descritas nas Seções a seguir. Na Seção 3.1, o pré-processamento dos dados. Na Seção 3.2, a mineração de padrões frequentes. E na Seção 3.3, a experimentação do dataset na tarefa de classificação utilizando técnicas de aprendizado de máquina.

3.1 Pré-processamento

Os dados provenientes do dataset possuem ruídos e muitas inconsistências em relação as possíveis classes dos atributos. Assim, uma limpeza dos dados foi realizada utilizando como referência o *Secondary Mushroom*, convertendo dados fora do escopo para nulos. Em seguida, os valores nulos (já existentes e os novos convertidos) foram preenchidos de acordo com as características de cada atributo. Os atributos contínuos passaram por uma detecção de *outliers* e normalização utilizando *z-score*. Os detalhes para as etapas de limpeza, preenchimento e padronização estão disponíveis no notebook da implementação[5]. A Figura 3 apresenta a distribuição dos atributos da partição de treino a partir dessas etapas de pré-processamento.

3.2 Mineração de Padrões Frequentes

3.2.1 Regras de Associações. Foram utilizados algoritmos de mineração de dados para descobrir padrões frequentes que aparecem na distribuição de características dos cogumelos. Para essa análise, os atributos contínuos foram discretizados (Figura 4). Em seguida, o dataset foi convertido para um formato binário, onde todos atributos possuem apenas valores verdadeiros e falsos, relativos as classes dos atributos originais.

Para a mineração de padrões frequentes, foi utilizado o algoritmo Apriori [1]. Com os seguintes parâmetros: $support \geq 0.05$, $confidence \geq 0.70$, $lift \geq 1.00$ e $conviction \geq 1.00$, sendo encontradas 66,566 regras com 1,460 consequentes únicos. A Tabela 4 apresenta as 5 regras de associação mineradas ordenadas por *lift*. O mesmo foi realizado analisando como consequente as classes da variável objetivo (comestível ou venenoso). Para consequente comestível, os parâmetros foram: $support \geq 0.015$, $confidence \geq 0.995$, $lift \geq 1.00$ e $conviction \geq 1.00$, sendo encontradas 15 regras (Tabela 5). Já para consequente venenoso: $support \geq 0.025$, $confidence \geq 0.90$, $lift \geq 1.00$ e $conviction \geq 1.00$, sendo encontradas 76 regras (Tabela 6). Mais detalhes são disponibilizados no notebook da implementação[6].

3.2.2 Sequências Frequentes. Para criar uma sequência com o dataset, foi utilizado características das dimensões dos cogumelos. Como existem três características que nos dão uma ideia da grandeza do cogumelo em um espaço tridimensional, calculamos seu volume estimado a partir dessas grandezas: `cap-diameter`, `stem-height` e `stem-width`. Uma sequência é criada para observar a mudança dos chápeus dos cogumelos de acordo com seu crescimento, ordenado pelo novo atributo de volume. Agrupamos essa variação do chapéu em função do volume utilizando outras características morfológicas disponíveis, criando mais de 2.000 sequências a serem analisadas pelo algoritmo PrefixSpan [4] com $support \geq 0.05$. Com base nos resultados, é observado que cogumelos tendem a possuir uma superfície do chapeu (`cap-surface`) mais lisa (*smooth*) conforme seu tamanho aumenta. Já em relação ao formato do chapéu

[4] https://kaggle.com/competitions/playground-series-s4e8

[5] https://github.com/Victorgonl/PoisonousMushrooms/blob/main/notebooks/1-preprocessing.ipynb

[6] https://github.com/Victorgonl/PoisonousMushrooms/blob/main/notebooks/2-patterns.ipynb

Tabela 2: Atributos e seus possíveis valores para os cogumelos rotulados do *Secondary Mushroom*.

Atributo	Tipo	Valores
class	Binário	poisonous=p, edible=e
cap-diameter	Contínuo	Número real em cm
cap-shape	Nominal	bell=b, conical=c, convex=x, flat=f, sunken=s, spherical=p, others=o
cap-surface	Nominal	fibrous=i, grooves=g, scaly=y, smooth=s, shiny=h, leathery=l, silky=k, sticky=t, wrinkled=w, fleshy=e
cap-color	Nominal	brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k
does-bruise-bleed	Nominal	bruises-or-bleeding=t, no=f
gill-attachment	Nominal	adnate=a, adnexed=x, decurrent=d, free=e, sinuate=s, pores=p, none=f, unknown=?
gill-spacing	Nominal	close=c, distant=d, none=f
gill-color	Nominal	Mesmo que cap-color + none=f
stem-height	Contínuo	Número real em cm
stem-width	Contínuo	Número real em mm
stem-root	Nominal	bulbous=b, swollen=s, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r
stem-surface	Nominal	Mesmo que cap-surface + none=f
stem-color	Nominal	Mesmo que cap-color + none=f
veil-type	Nominal	partial=p, universal=u
veil-color	Nominal	Mesmo que cap-color + none=f
has-ring	Binário	ring=t, none=f
ring-type	Nominal	cobwebby=c, evanescent=e, flaring=r, grooved=g, large=l, pendant=p, sheathing=s, zone=z, scaly=y, movable=m, none=f, unknown=?
spore-print-color	Nominal	Mesmo que cap-color
habitat	Nominal	grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d
season	Nominal	spring=s, summer=u, autumn=a, winter=w

Tabela 3: Distribuição de treino e teste para o *Kaggle Mushroom Dataset*. Os valores para o tipo de cogumelo não estão disponíveis abertamente para a partição de teste.

Partição	Instâncias	Porcentagem	Comestíveis	Venenosos
Treino	3.116.945	60%	1.411.549 (45%)	1.705.396 (55%)
Teste	2.077.964	40%	?	?

(cap-shape), a tendência é seu formato convexo (*convex*) se tornar mais achatado (*flat*) com o aumento de volume.

3.3 Avaliação Experimental

Com o pré-processamento do dados, é possível induzir um classificador a fim de predizer se um cogumelo é comestível ou venenoso. O classificador escolhido foi o CatBoost [5], um algoritmo de *gradient boosting* em árvores de decisão. A escolha em utilizar esse modelo é motivado pela natureza dos dados, sendo necessário suportar uma grande quantidade de dados sem a ocorrência de *overfitting*. Além disso, ter uma boa performance em dados categóricos e dados desbalanceados. O algoritmo também é eficiente em máquinas com recursos reduzidos, além de existir otimização por GPU.

Os experimentos seguiram em duas abordagens. A primeira (Seção 3.3.2), divide os dados entre treino, validação e teste. A segunda (Seção 3.3.3), utiliza 10 partições para realizar uma validação cruzada em 10 partes (*10-fold cross-validation*).

Durante ambas as abordagens, foi realizado a busca pela melhor combinação de hiper-parâmetros utilizando o *framework* Optuna [2]. A Tabela 7 apresenta o escopo de hiper-parâmetros a serem buscados durante as fases de validação. A métrica utilizada na busca é a mesma utilizada na avaliação das soluções e do desafio na plataforma Kaggle (Seção 3.3.1). Mais detalhes estão disponíveis no notebook da implementação[7].

3.3.1 Métricas de Avaliação. A métrica principal utilizada foi a MCC (*Matthews Correlation Coefficient*), sendo também utilizada no desafio do Kaggle. É uma medida utilizada para avaliar a qualidade de classificadores binários, levando em consideração todos os quatro valores da matriz de confusão: verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). É especialmente útil quando as classes estão desbalanceadas. A principal vantagem da MCC sobre outras métricas, como acurácia,

[7] https://github.com/Victorgonl/PoisonousMushrooms/blob/main/notebooks/3-experiments.ipynb

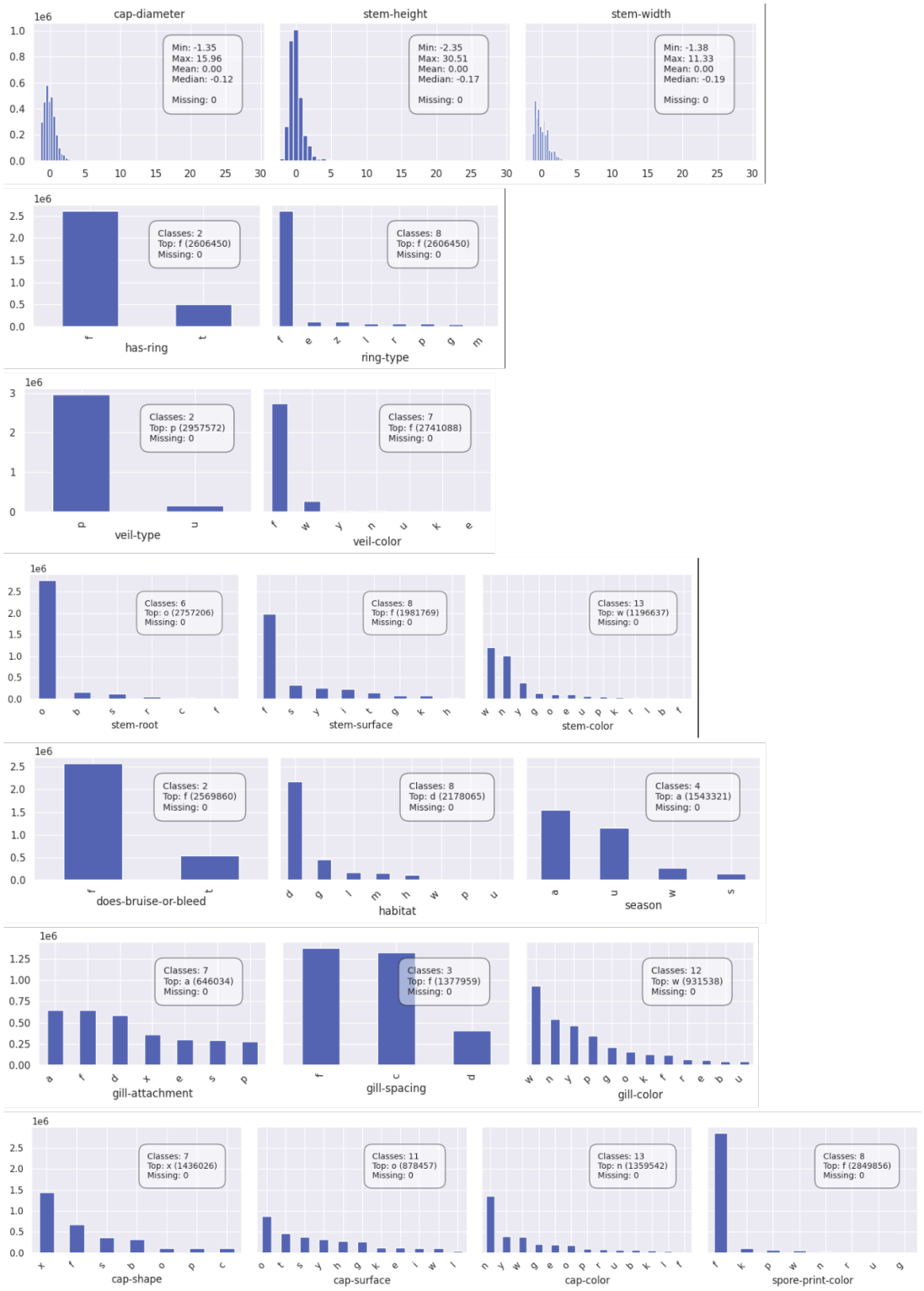


Figura 3: Distribuição dos atributos do dataset de treino pré-processado.

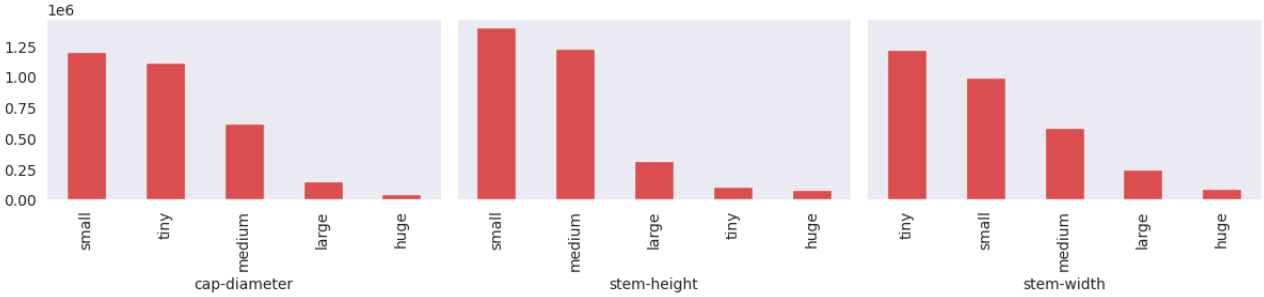


Figura 4: Discretização dos atributos contínuos em 5 classes.

Tabela 4: Top 5 regras de associações mineradas ordenadas por *lift*.

Antecedente	Consequente	Lift
Gill-attachment=free, gill-spacing=close, target=edible, cap-color=white	Ring-type=evanescent, stem-width=small, stem-color=white	65.673
Ring-type=evanescent, stem-color=white, target=edible	Gill-attachment=free, stem-width=small, gill-spacing=close, cap-color=white	59.264
Cap-surface=scaly, stem-width=tiny, gill-attachment=none	Cap-color=brown, cap-diameter=tiny, cap-shape=bell, stem-height=medium, gill-spacing=close, stem-color=white	59.112
Cap-color=brown, cap-diameter=tiny, cap-shape=bell, stem-height=medium, gill-spacing=close, stem-color=white	Cap-surface=scaly, stem-width=tiny, gill-attachment=none	59.112
Cap-color=brown, cap-shape=bell, stem-height=medium, gill-spacing=close, stem-width=tiny, stem-color=white	Cap-surface=scaly, cap-diameter=tiny, gill-attachment=none	58.826

Tabela 5: Top 5 regras de associação ordenadas por *lift* para cogumelos comestíveis

Antecedente	Consequente	Lift
Gill-attachment=free, stem-width=medium, stem-height=large, stem-color=white, does-bruise-or-bleed=yes	Edible	2.199
Gill-spacing=none, gill-attachment=free, stem-width=medium, stem-height=large, does-bruise-or-bleed=yes	Edible	2.199
Stem-height=large, gill-attachment=free, stem-width=medium, does-bruise-or-bleed=yes	Edible	2.199
Gill-spacing=none, gill-attachment=free, stem-height=large, stem-color=white, does-bruise-or-bleed=yes	Edible	2.199
Gill-attachment=free, stem-color=white, stem-height=large, does-bruise-or-bleed=yes	Edible	2.199

é que ela considera o equilíbrio entre as classes e penaliza erros desbalanceados. Além da MCC, também foi calculado as métricas acurácia, macro precisão, macro revocação e macro F-Score.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

O valor da MCC varia entre -1 e 1, sendo que:

- 1 indica uma predição perfeita;
- 0 indica uma predição aleatória;
- -1 indica uma total discordância entre predições e valores reais.

Tabela 6: Top 5 regras de associação ordenadas por *lift* para cogumelos venenosos

Antecedente	Consequente	Lift
Cap-color=brown, gill-spacing=none, stem-color=brown, stem-height=small, gill-color=brown, stem-width=tiny	Poisonous	1.822
Cap-color=brown, gill-spacing=none, stem-color=brown, gill-color=brown, stem-width=tiny	Poisonous	1.822
Gill-spacing=close, ring-type=others	Poisonous	1.821
Gill-spacing=none, stem-color=brown, gill-color=brown, stem-height=small, stem-width=tiny	Poisonous	1.821
Gill-spacing=none, gill-color=brown, stem-width=tiny, stem-color=brown	Poisonous	1.821

Tabela 7: Espaço de busca dos hiperparâmetros para otimização com Optuna e CatBoost

Parâmetro	Intervalo / Valor	Descrição
`iterations`	{500, 750, 1000}	Número máximo de árvores a serem treinadas.
`learning_rate`	[0.001, 0.01] (escala log)	Taxa de aprendizado do modelo.
`depth`	Inteiro de 4 a 10	Profundidade máxima das árvores.
`l2_leaf_reg`	Inteiro de 2 a 10	Coeficiente de regularização L2 nas folhas.
`border_count`	{32, 64, 128}	Número de divisões para binarização de atributos numéricos.
`loss_function`	`Logloss` (fixo)	Função de perda utilizada no treinamento.
`feature_border_type`	`GreedyLogSum` (fixo)	Estratégia para definição dos limites dos atributos contínuos.
`early_stopping_rounds`	50 (fixo)	Parada antecipada após iterações sem melhoria.

3.3.2 Experimento em Hold-Out. Para o experimento em *hold-out*, foi utilizado as primeiras sete partições como treino, a oitava como validação e as duas útimas como teste (70/10/20) . A fase de validação com busca de melhor combinação de hiper-parâmetros foi realizadas em 100 tentativas. A melhor combinação foi encontrada na tentativa número 76 obtendo um valor $MCC = 0.98199$, e é apresentada na Tabela 8. A Figura 5 apresenta a evolução da busca de hiper-parâmetros em função do número de tentativas. Em seguida, o modelo foi treinado novamente com essa melhor combinação, sendo testado no conjunto de teste (20%). Os resultados são apresentados na Tabela 9.

3.3.3 Experimento em 10-Fold Cross-Validation. O experimento de validação cruzada foi realizado seguindo o Algoritmo 1. A fase de validação com busca de hiper-parâmetros foi realizado em 25 tentativas para cada uma das 10 iterações. O resultado final é calculado com base na média aritmética dos resultados de cada uma das iterações junto ao desvio-padrão, apresentados na Tabela 10.

3.3.4 Desafio no Kaggle. Ao fim dos experimentos locais, dois modelos finais foram separados para inferir os dados de teste do *Kaggle Mushroom Dataset*, que, em seguida, foram submetidos à plataforma para avaliação. O primeiro modelo é treinado com todos os dados de treino do dataset da plataforma utilizando os hiper-parâmetros encontrados na Seção 3.3.2. Já o segundo é um classificador por votação (*Voting Classifier*) utilizando os 10 modelos gerados pelo experimento descrito na Seção 3.3.3, em ambas as modalidades *hard* e *soft*. Os resultados retornados pela plataforma são apresentados na Figura 6. Para comparação dos melhores resultados obtidos por participantes do desafio, a Figura 7 apresenta os top 10 resultados classificados.

Algorithm 1 Validação Cruzada em K partes com Busca de Hiper-Parâmetros.

1: Seja o dataset dividido em K partições.
2: **for** $i = 1$ **to** K **do**:
3: `teste` $\leftarrow k_i$.
4: `validação` $\leftarrow k_{(i+1 \bmod K)}$.
5: `treino` $\leftarrow$ restante das partições.
6: Busca melhor combinação de hiper-parâmetros usando `treino` e `validação`.
7: Avalia o melhor modelo encontrado com `teste` e salva as métricas.
8: **end for**
9: Calcula a média e o desvio-padrão das métricas salvas.

4 Conclusões

Durante o trabalho, foi abordado a classificação de cogumelos entre comestíveis e venenosos aplicando técnicas de Mineração de Dados como o pré-processamento e a busca de padrões frequentes. A tarefa de classificação foi implementada com o modelo CatBoost, capaz de lidar com o grande número de instâncias do dataset utilizado, sendo avaliado nas modalidades *hold-out* e *10-fold cross-validation* com a busca de melhor combinação de hiper-parâmetros. Os resultados demonstram que as etapas abordadas resultam em um classificador robusto, capaz de predizer com precisão se um cogumelo é veneno

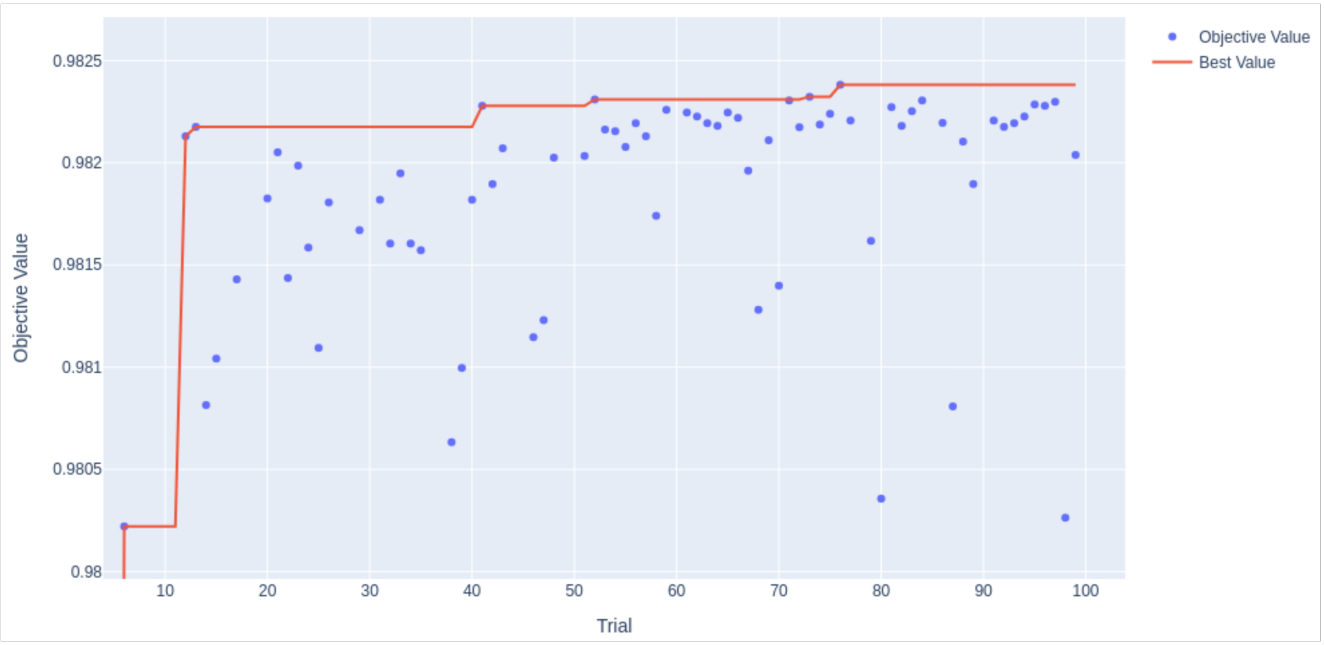


Figura 5: Evolução dos valores objetivos durante a busca de melhores hiper-parâmetros na fase de validação do experimento *hold-out*. O gráfico foi filtrado para apresentar apenas valores obtidos maiores que 0.98.

Tabela 8: Melhor combinação de hiper-parâmetros encontrada durante o experimento de *hold-out*.

		Hiper-Parâmetro				
Tentativa	**MCC**	`iterations`	`learning_rate`	`depth`	`l2_leaf_reg`	`border_count`
76	0.98238	1000	0.07320	12	7	64

Tabela 9: Resultados do experimento de *hold-out*.

Accuracy	Macro Precision	Macro Recall	Macro F-Score	MCC
0.99107	0.99091	0.99109	0.99010	0.98199

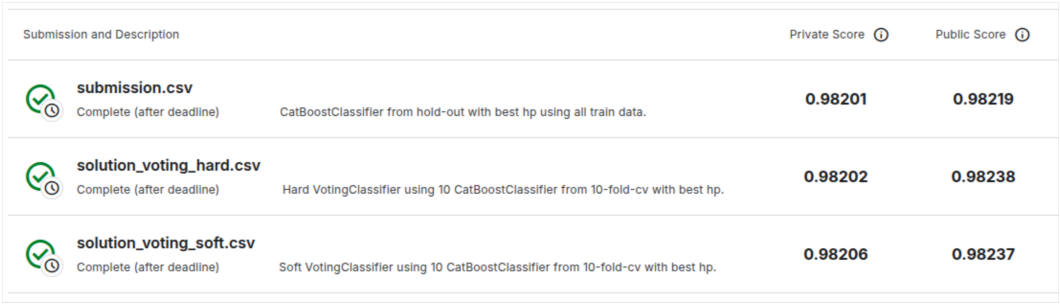
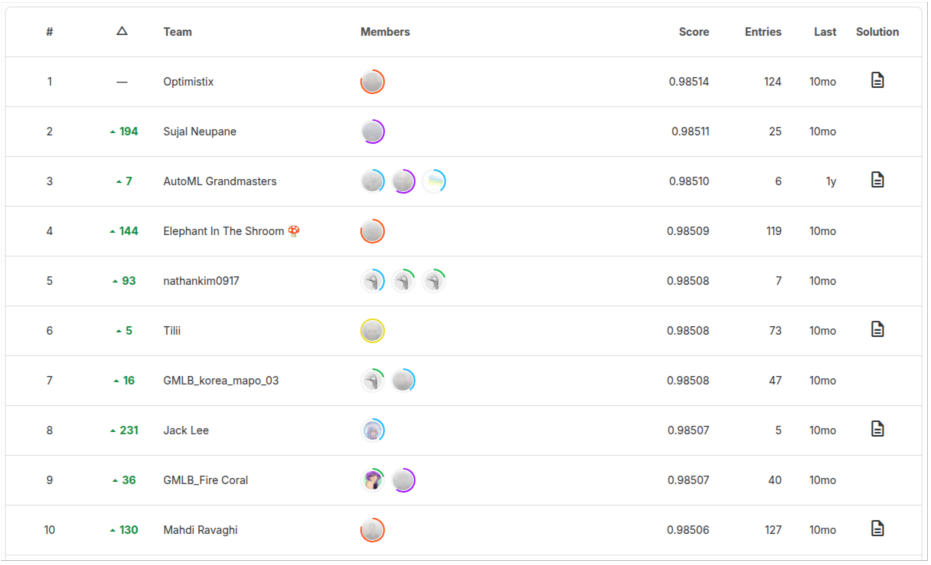


Figura 6: Resultados obtidos no desafio *Binary Prediction of Poisonous Mushrooms* da plataforma Kaggle. *Public* é calculado em 20% dos dados de teste e *private* em 80%, sendo este último o *benchmark* principal do desafio.

Tabela 10: Resultados do experimento de validação cruzada em 10 partes.

Fold	Accuracy	Precision	Recall	F1-score	MCC
1	0.99081	0.99212	0.99109	0.99160	0.98147
2	0.99085	0.99286	0.99039	0.99162	0.98155
3	0.99066	0.99232	0.99054	0.99143	0.98117
4	0.99115	0.99268	0.99117	0.99192	0.98213
5	0.99117	0.99292	0.99092	0.99192	0.98218
6	0.99077	0.99271	0.99041	0.99156	0.98138
7	0.99072	0.99241	0.99062	0.99151	0.98127
8	0.99095	0.99269	0.99076	0.99173	0.98173
9	0.99109	0.99282	0.99086	0.99184	0.98203
10	0.99108	0.99275	0.99094	0.99184	0.98200
$\bar{x} \pm \sigma$	0.99092 ± 0.00018	0.99263 ± 0.00025	0.99077 ± 0.00026	0.99170 ± 0.00017	0.98169 ± 0.00036



#	△	Team	Members	Score	Entries	Last	Solution
1	—	Optimistix		0.98514	124	10mo	
2	▲ 194	Sujal Neupane		0.98511	25	10mo	
3	▲ 7	AutoML Grandmasters		0.98510	6	1y	
4	▲ 144	Elephant In The Shroom 🍄		0.98509	119	10mo	
5	▲ 93	nathankim0917		0.98508	7	10mo	
6	▲ 5	Tilii		0.98508	73	10mo	
7	▲ 16	GMLB_korea_mapo_03		0.98508	47	10mo	
8	▲ 231	Jack Lee		0.98507	5	10mo	
9	▲ 36	GMLB_Fire Coral		0.98507	40	10mo	
10	▲ 130	Mahdi Ravaghi		0.98506	127	10mo	

Figura 7: Resultados obtidos (*private*) no desafio *Binary Prediction of Poisonous Mushrooms* da plataforma Kaggle.

ou não. As predições do conjunto de teste foram submetidos a plataforma Kaggle, obtendo valores para a métrica MCC entre 0.98201 e 0.98206. Para trabalhos futuros, é interessante abordar a classificação por votação utilizando um número maior de modelos variados, assim como são abordados por outras soluções que aparecem no ranking do desafio.

Referências

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, page 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1558601538.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019.

[3] Wacharaphol Ketwongsa, Sophon Boonlue, and Urachart Kokaew. A new deep learning model for the classification of poisonous and edible mushrooms based on improved alexnet convolutional neural network. *Applied Sciences*, 12(7), 2022. ISSN 2076-3417.

[4] Jian Pei, Jiawei Han, B. Mortazavi-Asl, H. Pinto, Qiming Chen, U. Dayal, and Mei-Chun Hsu. Prefixspan,: mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings 17th International Conference on Data Engineering*, pages 215–224, 2001.

[5] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6639–6649, Red Hook, NY, USA, 2018. Curran Associates Inc.

[6] Kemal Tutuncu, Ilkay Cinar, Ramazan Kursun, and Murat Koklu. Edible and poisonous mushrooms classification by machine learning algorithms. In *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–4, 2022.

[7] Kemal Tutuncu, Ilkay Cinar, Ramazan Kursun, and Murat Koklu. Research on mushroom classification based on xgb technology. In *2025 International Conference on Biomaterials, Computer Engineering and Ecological Environment*, volume 13, 2025.

[8] Dennis Wagner, Dominik Heider, and Georges Hattab. Mushroom data creation, curation, and simulation to support classification tasks. *Scientific Reports*, 11, 2021.