Análise Preditiva da Comestibilidade de Cogumelos com Algoritmos de Aprendizado de Máquina

Bruno Castro Tomaz[1], Tomás Fiorelli Barbosa[1]

[1]Faculdade de Computação e Informática
Universidade Presbiteriana Mackenzie (UPM)
Rua da Consolação, 930 – Higienópolis – 01302-907 – São Paulo, SP – Brasil

{10389988,10395687}@mackenzista.com.br

Resumo. *Este relatório investiga o uso de algoritmos de aprendizado de máquina para classificar cogumelos como comestíveis ou venenosos, utilizando um dataset público com características físicas e ecológicas. Foram testados sete algoritmos: KNN, Decision Tree, Logistic Regression, Naive Bayes, SVM, Random Forest e AdaBoost. Modelos baseados em árvores e proximidade, como Random Forest e KNN, alcançaram acurácias superiores a 99%, destacando-se pela eficácia. Modelos probabilísticos e lineares, como Naive Bayes e Logistic Regression, apresentaram desempenhos inferiores. Os resultados confirmam que a escolha dos algoritmos e a preparação adequada dos dados são fundamentais para classificações precisas e confiáveis.*

1. Introdução

a. Contextualização

O consumo de cogumelos é amplamente difundido em diversas culturas, tanto por seu valor nutricional quanto por seu uso em práticas medicinais. No entanto, a identificação de cogumelos comestíveis e venenosos é um desafio crítico, especialmente para pessoas sem conhecimento especializado. Erros nessa identificação podem levar a consequências graves, incluindo intoxicação alimentar e até a morte.

Com o avanço das tecnologias de aprendizado de máquina, tornou-se possível abordar esse problema utilizando modelos preditivos para identificar automaticamente a comestibilidade de cogumelos com base em suas características. A aplicação desses modelos não apenas aumenta a segurança, mas também contribui para a conscientização sobre o consumo responsável e o uso sustentável de recursos naturais.

b. Justificativa

A escolha desse tema justifica-se pela relevância da segurança alimentar e pelo impacto que soluções tecnológicas podem trazer para a identificação de alimentos seguros. Além disso, a classificação de cogumelos é um caso de uso ideal para demonstrar o potencial de algoritmos de aprendizado de máquina supervisionado em resolver problemas de classificação binária. Por meio dessa abordagem, espera-se explorar o uso de ferramentas modernas para obter insights a partir de dados e, ao mesmo tempo, contribuir para a prevenção de riscos associados ao consumo inadequado de cogumelos.

c. Objetivo

O objetivo deste trabalho é aplicar algoritmos de aprendizado de máquina supervisionado do tipo classificação, fazendo uso do *dataset* de cogumelos Secondary Data, um problema clássico utilizado nessa área de pesquisa, e gerar um modelo preditivo que permita classificá-los nas categorias comestíveis ou venenosos, com boa assertividade.

Na sua geração, pretende-se realizar experimentos com os algoritmos de classificação mais adequados para problemas dessa natureza, incluindo Árvore de Decisão, Random Forest e AdaBoost, avaliando seus desempenhos com base em métricas como acurácia, precisão e recall. Esses algoritmos foram selecionados por sua eficiência e ampla aplicação em problemas de classificação complexos.

d. Opção de projeto

Este projeto se enquadra na "Opção Framework", que envolve o uso de um framework ou ferramenta de aprendizado de máquina, como o scikit-learn, para solucionar um problema de classificação. A escolha desse formato permite explorar algoritmos supervisionados para abordar o desafio de distinguir cogumelos comestíveis de venenosos, utilizando um conjunto de dados pré-existente.

2. Descrição do Problema

A classificação de cogumelos em comestíveis ou venenosos é um problema relevante e desafiador devido à diversidade de espécies e características físicas, muitas vezes sutis, que podem diferenciar os tipos seguros dos perigosos. Para pessoas sem conhecimento especializado, essa tarefa pode ser quase impossível, aumentando o risco de consumo acidental de espécies tóxicas, que podem causar desde desconforto gastrointestinal até óbito.

Do ponto de vista técnico, o problema pode ser formulado como uma tarefa de classificação binária, onde o objetivo é prever a comestibilidade (ou toxicidade) de um cogumelo com base em suas características morfológicas e físicas, como cor, forma do chapéu, textura, odor e outras propriedades. Os dados utilizados são provenientes de um dataset público que contém registros detalhados de cogumelos, incluindo informações sobre quais são comestíveis e quais são venenosos.

Esse problema apresenta desafios típicos de aprendizado de máquina, como:

- **A complexidade das variáveis**: Muitas das características disponíveis são categóricas e precisam ser adequadamente codificadas para serem interpretadas pelos algoritmos.
- **Relevância de atributos**: Nem todas as variáveis disponíveis contribuem de forma significativa para a previsão, sendo necessário identificar quais têm maior impacto no modelo.
- **Generalização**: Garantir que o modelo desenvolvido tenha um bom desempenho não apenas no conjunto de treinamento, mas também em novos dados, evitando problemas como *overfitting*.

A resolução desse problema não só demonstra o potencial de algoritmos de aprendizado de máquina em aplicações práticas, mas também contribui para a segurança alimentar e a conscientização sobre o consumo responsável de cogumelos.

3. *Dataset*

O *dataset* utilizado neste trabalho é composto por aproximadamente 61 mil registros simulados, baseados nas descrições de 173 espécies reais de cogumelos. Essas descrições incluem informações detalhadas sobre 22 categorias distintas, representando características morfológicas e ambientais dos cogumelos. A variável alvo da análise é **class** que indica se um cogumelo é **edible** (*e*, comestível) ou **poisonous** (*p*, venenoso). O objetivo é realizar a classificação binária com base nos atributos disponíveis, permitindo identificar com precisão a comestibilidade dos cogumelos. Os atributos do *dataset* vieram organizados da seguinte forma:

- **cap-diameter (m):** Diâmetro do chapéu (float em cm).
- **cap-shape (n):** Formato do chapéu, com categorias: bell (*b*), conical (*c*), convex (*x*), flat (*f*), sunken (*s*), spherical (*p*), others (*o*).
- **cap-surface (n):** Superfície do chapéu, com categorias: fibrous (*i*), grooves (*g*), scaly (*y*), smooth (*s*), dry (*d*), shiny (*h*), leathery (*l*), silky (*k*), sticky (*t*), wrinkled (*w*), fleshy (*e*).
- **cap-color (n):** Cor do chapéu, com categorias: 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 (n):** Indica se o cogumelo machuca ou sangra: bruises-or-bleeding (*t*), no (*f*).
- **gill-attachment (n):** Tipo de fixação das lamelas, com categorias: adnate (*a*), adnexed (*x*), decurrent (*d*), free (*e*), sinuate (*s*), pores (*p*), none (*f*), unknown (*?*).
- **gill-spacing (n):** Espaçamento das lamelas: close (*c*), distant (*d*), none (*f*).
- **gill-color (n):** Cor das lamelas (veja **cap-color**) + none (*f*).
- **stem-height (m):** Altura do caule (float em cm).
- **stem-width (m):** Largura do caule (float em mm).
- **stem-root (n):** Tipo de raiz do caule, com categorias: bulbous (*b*), swollen (*s*), club (*c*), cup (*u*), equal (*e*), rhizomorphs (*z*), rooted (*r*).
- **stem-surface (n):** Superfície do caule (veja **cap-surface**) + none (*f*).
- **stem-color (n):** Cor do caule (veja **cap-color**) + none (*f*).
- **veil-type (n):** Tipo de véu: partial (*p*), universal (*u*).
- **veil-color (n):** Cor do véu (veja **cap-color**) + none (*f*).
- **has-ring (n):** Indica se o cogumelo possui anel: ring (*t*), none (*f*).
- **ring-type (n):** Tipo de anel, com categorias: 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 (n):** Cor da impressão de esporos (veja **cap-color**).

- **habitat (n):** Habitat natural do cogumelo, com categorias: grasses (*g*), leaves (*l*), meadows (*m*), paths (*p*), heaths (*h*), urban (*u*), waste (*w*), woods (*d*).
- **season (n):** Estação do ano: spring (*s*), summer (*u*), autumn (*a*), winter (*w*).
- **class (n):** Classe do cogumelo, objetivo da classificação: edible (*e*), poisonous (*p*).

Esse conjunto de atributos oferece uma ampla gama de informações para a construção de modelos preditivos, permitindo a exploração de diferentes algoritmos de aprendizado de máquina na tarefa de classificação. O uso de atributos simulados, mas baseados em espécies reais, proporciona um cenário realista e escalável para experimentação.

3.1. Análise Exploratória

Nesta etapa do projeto, realizamos a análise exploratória e a preparação dos dados utilizando o ambiente Jupyter Notebook. A análise foi conduzida com o objetivo de identificar padrões e preparar os dados para os modelos preditivos.

Inicialmente, o *dataset* foi importado e examinamos sua estrutura, incluindo o número de registros, colunas e tipos de variáveis. A presença de valores ausentes foi identificada e, como estratégia, as colunas com mais de 40% de valores ausentes foram removidas. Para os valores ausentes restantes, foi aplicada a imputação com a string "NO_DATA", permitindo manter a consistência no conjunto de dados.

Posteriormente, realizamos uma análise descritiva das variáveis, gerando estatísticas básicas e visualizações como histogramas e gráficos de barras para as variáveis categóricas. Com isso, passamos de 21 atributos para 15, incluindo a variável alvo, com o mesmo total de registros, 61.069, conforme mostra a descrição abaixo:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61069 entries, 0 to 61068
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   class                 61069 non-null  object
 1   cap-diameter          61069 non-null  float64
 2   cap-shape             61069 non-null  object
 3   cap-surface           61069 non-null  object
 4   cap-color             61069 non-null  object
 5   does-bruise-or-bleed  61069 non-null  object
 6   gill-attachment       61069 non-null  object
 7   gill-color            61069 non-null  object
 8   stem-height           61069 non-null  float64
 9   stem-width            61069 non-null  float64
 10  stem-color            61069 non-null  object
 11  has-ring              61069 non-null  object
 12  ring-type             61069 non-null  object
 13  habitat               61069 non-null  object
 14  season                61069 non-null  object
dtypes: float64(3), object(12)
memory usage: 7.0+ MB
```

Figura 1 – Execução do comando *DataFrame.info()*

Para garantir que as variáveis categóricas fossem adequadamente interpretadas pelos algoritmos de aprendizado de máquina, aplicamos a técnica de *One-Hot Encoding*. Essa abordagem transformou as categorias em variáveis binárias, permitindo sua inclusão em modelos que requerem entradas numéricas.

A análise exploratória incluiu também a geração de matrizes de confusão para avaliar o desempenho inicial dos modelos preditivos. Esses gráficos forneceram insights valiosos sobre as classes mais frequentemente confundidas, além de indicar possíveis padrões nos erros de classificação.

A preparação dos dados foi fundamental para garantir um conjunto robusto e adequado para os modelos de aprendizado de máquina que foram utilizados nas etapas subsequentes. As transformações aplicadas foram baseadas em melhores práticas e critérios objetivos, assegurando uma base sólida para a construção de modelos preditivos confiáveis.

4. Metodologia

Nesta etapa do projeto, utilizaremos diferentes algoritmos de aprendizado de máquina para realizar a classificação da comestibilidade dos cogumelos. A escolha dos modelos foi feita de forma a explorar diversas abordagens de aprendizado supervisionado, incluindo métodos baseados em proximidade, regras, probabilidade, margens de separação e *ensembles*. Os algoritmos escolhidos — K-Nearest Neighbors (KNN), Decision Tree Classifier (CART), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RFor) e AdaBoost (AdaB) — apresentam características complementares, permitindo uma análise abrangente do desempenho preditivo. Cada um deles será avaliado em termos de acurácia, matriz de confusão e métricas adicionais, com o objetivo de identificar o modelo mais adequado para a tarefa em questão.

4.1. K-Nearest Neighbors (KNN)

O KNN é um algoritmo de aprendizado baseado em instâncias que classifica um dado ponto considerando a proximidade aos kkk vizinhos mais próximos no espaço das características. Ele é simples, intuitivo e funciona bem para conjuntos de dados menores ou bem distribuídos. A desvantagem está no custo computacional alto para grandes volumes de dados, já que precisa calcular distâncias para todas as instâncias.

4.2. Decision Tree Classifier (CART)

As árvores de decisão [1, 2] são algoritmos baseados em regras que dividem o espaço dos dados em regiões com base em condições lógicas (exemplo: "se X > valor, então classe Y"). Elas são interpretáveis e úteis para identificar relações entre características. Contudo, podem sofrer de *overfitting*, especialmente sem poda ou regularização.

4.3. Logistic Regression (LR)

Embora o nome sugira uma técnica de regressão, a regressão logística é usada para classificação binária. Ela modela a probabilidade de uma instância pertencer a uma classe usando uma função sigmoide e é eficaz em problemas lineares. Não lida bem com relações não lineares entre as variáveis, a menos que sejam introduzidas características derivadas.

4.4. Naive Bayes (NB)

O Naive Bayes é baseado no teorema de Bayes e assume que as características são condicionalmente independentes. Apesar dessa suposição simplificadora, ele é eficiente, rápido e funciona bem em conjuntos de dados textuais e categóricos. Pode ser menos eficaz quando as características são altamente correlacionadas.

4.5. Support Vector Machine (SVM)

As SVMs são algoritmos baseados em margens que tentam encontrar o hiperplano que melhor separa as classes no espaço das características. Elas funcionam bem em problemas com limites de decisão complexos e podem usar *kernels* para modelar relações não lineares. No entanto, são computacionalmente intensivas e podem ser sensíveis ao balanceamento de classes.

4.6. Random Forest (RFor)

O Random Forest [3] combina várias árvores de decisão independentes, criadas a partir de amostras do dataset, e realiza a classificação por votação da maioria. É um método robusto, resistente ao *overfitting* e capaz de lidar com dados de alta dimensionalidade. Sua desvantagem está no custo computacional em datasets muito grandes.

4.7. AdaBoost (AdaB)

O AdaBoost é um algoritmo de *ensemble* que combina vários classificadores fracos (geralmente árvores de decisão simples) para formar um modelo mais robusto. Ele ajusta iterativamente os pesos das instâncias mal classificadas, aumentando sua importância no treinamento subsequente. Pode ser sensível a ruídos e *outliers*.

5. Resultados

Os algoritmos utilizados neste projeto apresentaram desempenhos variados na classificação da comestibilidade dos cogumelos, conforme a tabela de acurácia abaixo:

	nome	acuracia
5	RFor	99.983625
0	KNN	99.860815
1	CART	99.762568
4	SVM	89.659407
6	AdaB	79.998363
2	LR	78.131652
3	NB	66.923203

Tabela 1 – Acurácia dos algoritmos de aprendizado de máquina

Os três modelos de maior desempenho foram o Random Forest (99,98%), o K-Nearest Neighbors (99,86%) e o Decision Tree Classifier (99,76%), que obtiveram resultados excepcionalmente altos, indicando excelente capacidade de aprendizado e classificação. Isso se deve à natureza dos dados e à alta relevância das variáveis preditoras

na separação entre cogumelos comestíveis e venenosos. Esses modelos, no entanto, foram analisados detalhadamente para garantir que o alto desempenho não fosse resultado de *overfitting*. A validação cruzada e a análise das matrizes de confusão reforçaram a robustez de seus resultados.

Os outros modelos apresentaram desempenhos significativamente inferiores. O Support Vector Machine (SVM) teve uma acurácia de 89,66%, mostrando-se adequado para o problema, embora tenha sido superado pelos métodos baseados em árvores e proximidade. O AdaBoost (79,99%) e a Logistic Regression (78,13%) apresentaram desempenho mediano, refletindo limitações em capturar as complexidades dos dados. Por fim, o Naive Bayes (66,92%) apresentou o menor desempenho, o que pode ser atribuído à sua suposição de independência entre as variáveis, que não se aplica bem ao dataset em questão.

Esses resultados indicam que métodos baseados em árvores e *ensemble* foram os mais eficazes, aproveitando sua capacidade de lidar com interações complexas entre as variáveis. No entanto, o desempenho geral dos modelos foi consistente com as características intrínsecas de cada algoritmo e a natureza do problema de classificação.

6. Conclusão

Os resultados obtidos neste estudo, conduzido para classificar a comestibilidade de cogumelos com base em suas características físicas e ecológicas, demonstram a eficácia do aprendizado de máquina na resolução de problemas de classificação em dados multivariados. A proposta inicial do projeto era identificar padrões que distinguissem cogumelos comestíveis dos venenosos, utilizando técnicas de aprendizado supervisionado para explorar relações entre as variáveis preditoras.

A análise revelou que modelos baseados em árvores, como Random Forest e Decision Tree, e métodos de proximidade, como KNN, alcançaram desempenho excepcional, com acurácia superior a 99%. Esses algoritmos mostraram-se altamente adequados para o problema, devido à sua capacidade de capturar interações complexas e explorar a riqueza das variáveis disponíveis. Por outro lado, modelos probabilísticos e lineares, como Naive Bayes e Logistic Regression, apresentaram limitações no tratamento das relações intrínsecas dos dados, mas ainda assim contribuíram para uma compreensão mais ampla do problema.

Em conclusão, o projeto evidenciou que a escolha de algoritmos robustos, combinada com uma preparação criteriosa dos dados, é fundamental para alcançar predições confiáveis em problemas de classificação. A abordagem multifacetada adotada reforça a importância de explorar diferentes técnicas para identificar as soluções mais adequadas, destacando o papel do aprendizado de máquina na construção de sistemas preditivos eficientes e aplicáveis a cenários práticos.

7. GitHub e Vídeo Explicativo

GitHub: https://github.com/tomasfiorelli/mushroom-analysis/tree/main

Vídeo: https://www.youtube.com/watch?v=8EKMiIveRTA

8. Referências

[1] HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd ed. Elsevier, 2011.

[2] MITCHELL, T. *Machine Learning*. 1st ed. McGraw-Hill, 1997.

[3] BREIMAN, L. *Random Forests*. Machine Learning, v.45, p. 5–32, 2001.

9. Bibliografia

ghattab. Secondary Data. Disponível em: https://github.com/ghattab/secondarydata/. Acesso em: 09/11/2024.