Trabalho 5

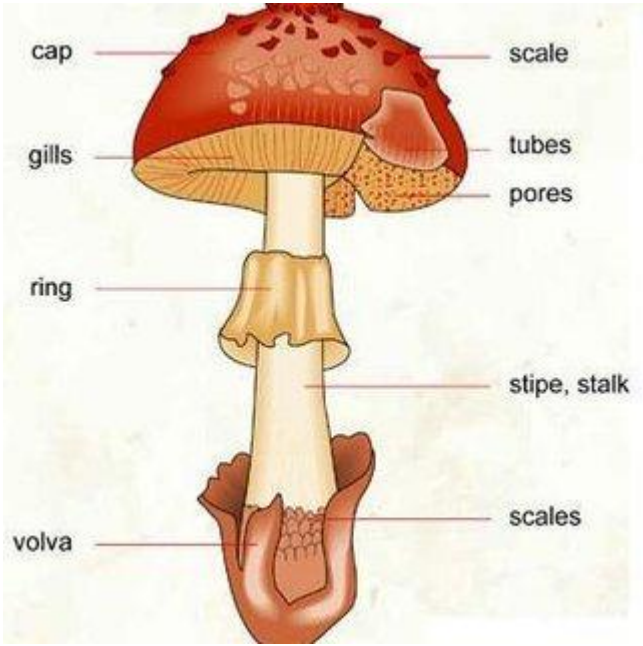
Classificação de cogumelos

Utilizando árvore de decisão

Jonathan Suhett Barbêdo
Renan Mendanha Alvarino

Base de Dados

- A base de dados escolhida classifica cogumelos entre venenosos e comestíveis
- São ao todo 8124 elementos
- Não há dados faltando
- Cada um dos cogumelos possui 22 atributos
- 52% da instancias comestíveis e 48% venenosos, caracterizando um bom balanceamento de dados.



Base de Dados

- cap_shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
- cap_surface: fibrous=f,grooves=g,scaly=y,smooth=s
- cap_color: brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y
- bruises: bruises=t,no=f
- odor: almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s
- gill_attachment: attached=a,descending=d,free=f,notched=n
- gill_spacing: close=c,crowded=w,distant=d
- gill_size: broad=b,narrow=n
- gill_color: black=k,brown=n,buff=b,chocolate=h,gray=g,
  green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y
- stalk_shape: enlarging=e,tapering=t
- stalk_root: bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?

Base de Dados

- stalk_surface_above_ring: fibrous=f,scaly=y,silky=k,smooth=s
- stalk_surface_below_ring: fibrous=f,scaly=y,silky=k,smooth=s
- stalk_color_above_ring:
  brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
- stalk_color_below_ring:
  brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
- veil_type: partial=p,universal=u
- veil_color: brown=n,orange=o,white=w,yellow=y
- ring_number: none=n,one=o,two=t
- ring_type: cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z
- spore_print_color:
  black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y
- population: abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y
- habitat: grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d

Base de Dados

- Número total de atributos criados na árvore = 114

- Número total de instâncias possíveis = 60,949,905,408,000

Atributo	tipos do atributo
cap_shape	6
cap_surface	4
cap_color	10
bruises	2
odor	9
gill_attachment	1 -> False (podemos eliminar esse atributo)
gill_spacing	2
gill_size	2
gill_color	12
stalk_shape	2
stalk_root	5
stalk_surface_above_ring	4
stalk_surface_below_ring	4
stalk_color_above_ring	9
stalk_color_below_ring	9
veil_type	1 -> p (podemos eliminar esse atributo)
veil_color	4
ring_number	3
ring_type	5
spore_print_color	9
population	6
habitat	7

Experimento base

Como não há dados incompletos, vamos utilizar nessa primeira etapa todos os atributos que não foram eliminados;

Para divisão treinamento/teste, utilizamos k-fold = 5;

A função utilizada para escolher o atributo é a Entropia;

Com esses parâmetros iniciais, obtivemos os seguintes resultados em relação ao conjunto de testes:

Experimento base

Acurácia:

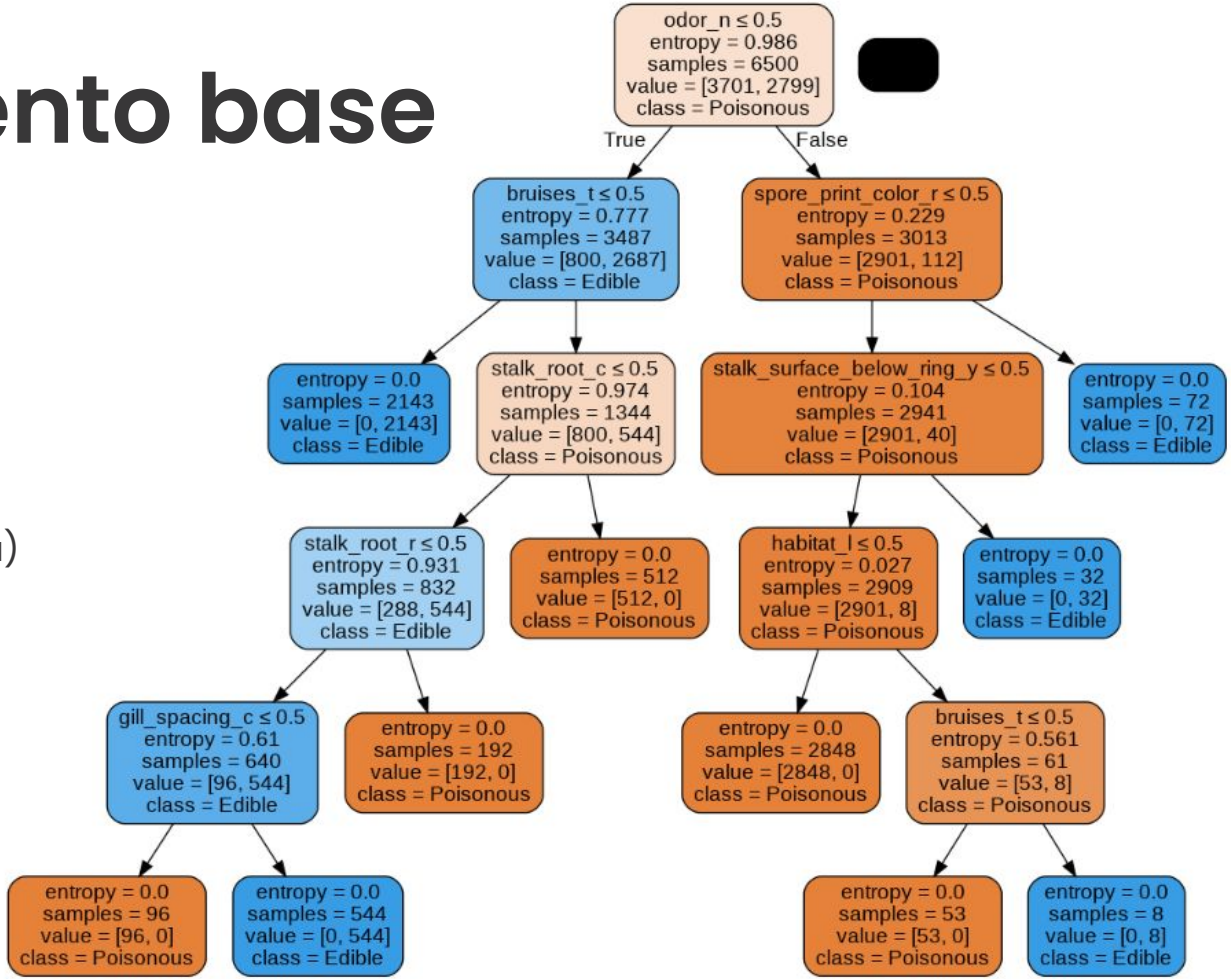
1. 1.0
2. 1.0
3. 1.0
4. 0.9926153846153846
5. 0.9901477832512315

0.9965526335733232 (média)

Precisão:

1. 1.0
2. 1.0
3. 1.0
4. 0.9955914768552535
5. 0.9929390997352162

0.997706115318094 (média)

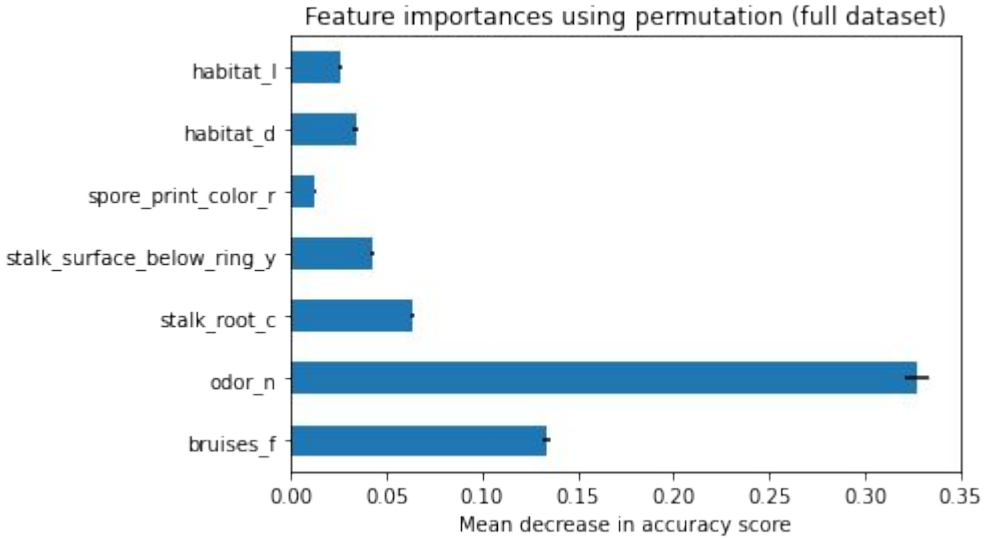


Experimento base

- Matriz com falsos positivos e falsos negativos (test set confusion matrix)

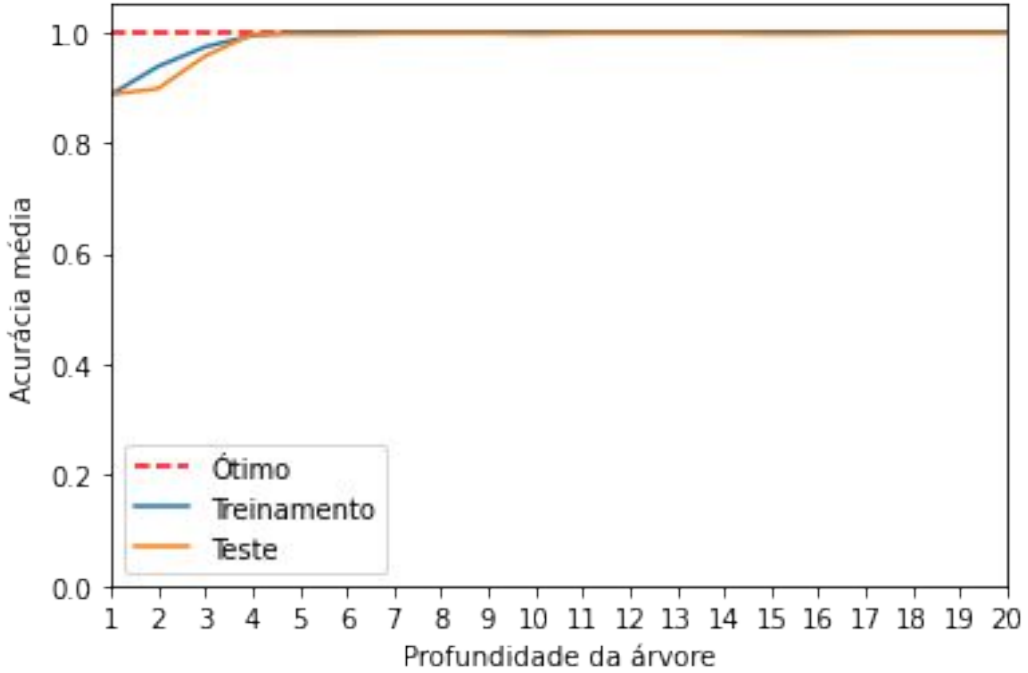
	Predicted edible	Predicted poisonous
Real edible	491	16
Real poisonous	0	1117

- Importância dos atributos



Experimento base

- Podas de profundidade da árvore em relação à acurácia média



Alterando número de folds

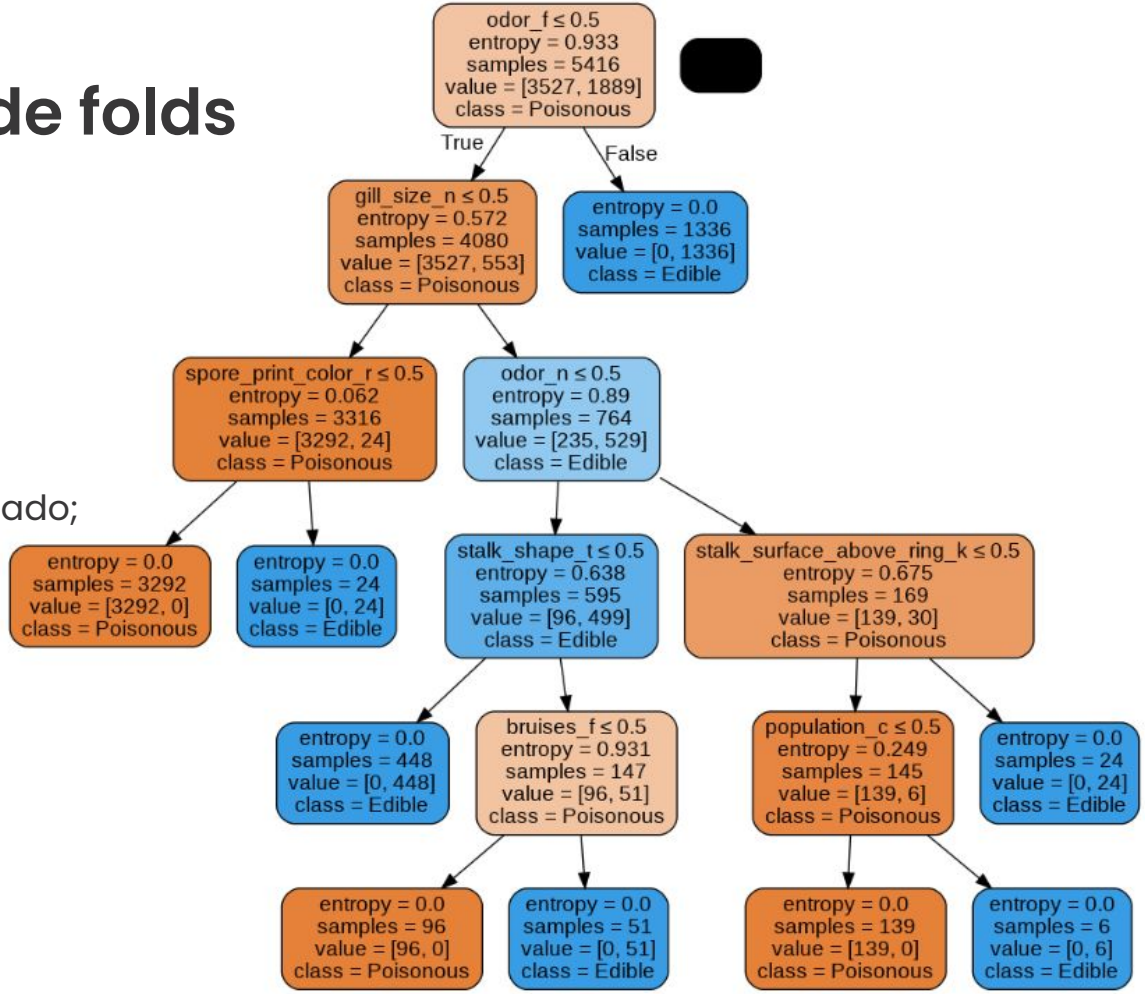
Para k-fold= 10, acurácia=1
em todos os folds;

Para k-fold=3, geramos a árvore ao lado;

E acurácia abaixo:

1. 0.7045790251107829
2. 1.0
3. 0.9867060561299852

0.897095027080256 (média)

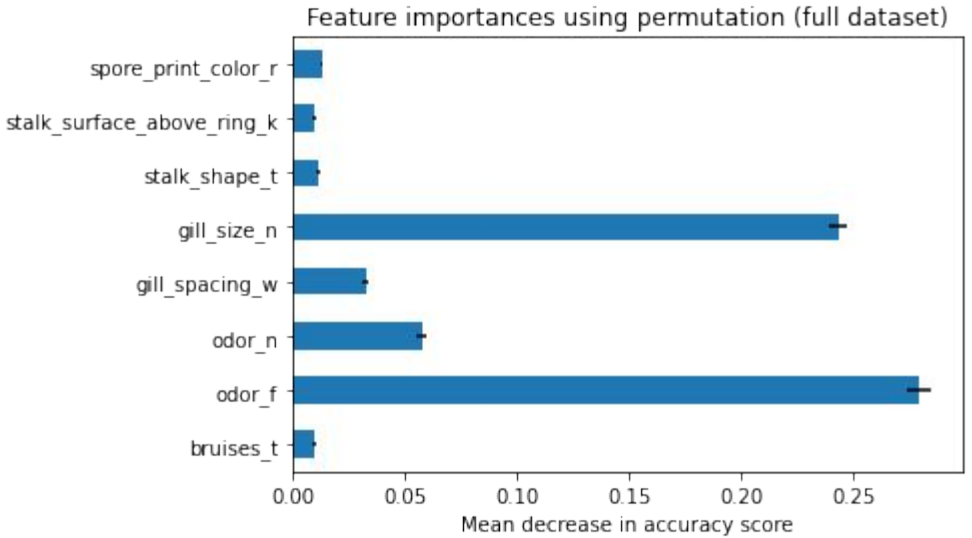


Alterando número de folds

- Matriz com falsos positivos e falsos negativos (test set confusion matrix)

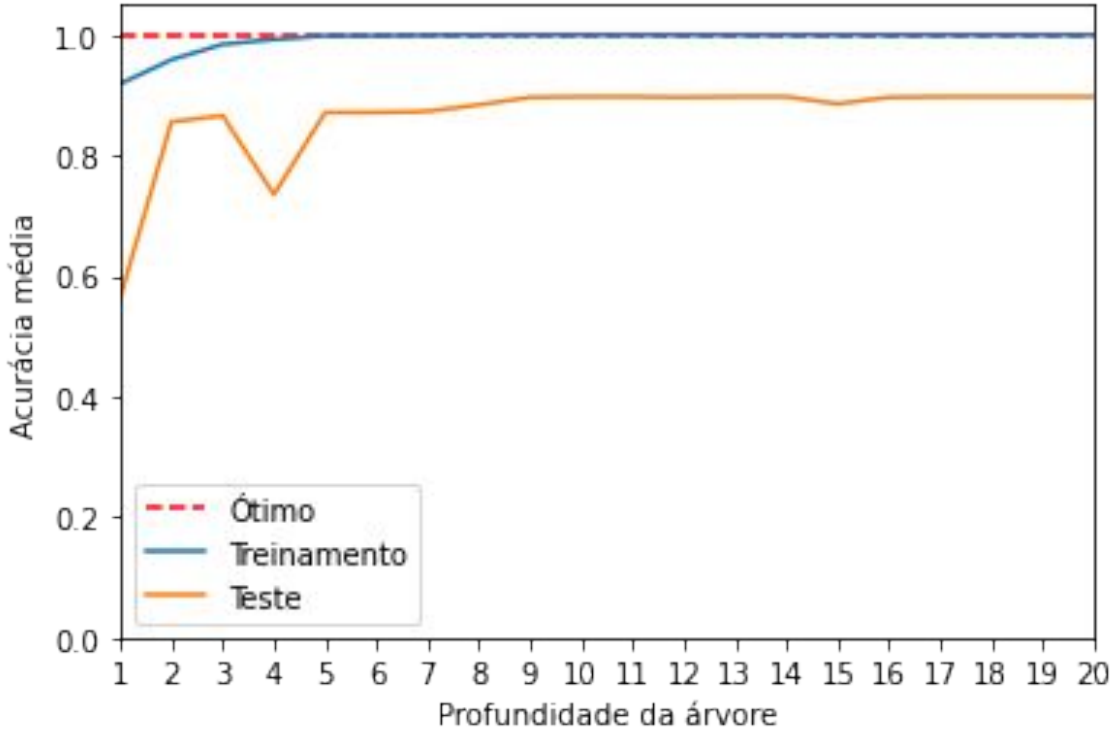
	Predicted edible	Predicted poisonous
Real edible	681	0
Real poisonous	44	1983

- Importância dos atributos



Alterando número de folds

- Podas de profundidade da árvore em relação à acurácia média



Alterando número de folds

Percebemos que o classificador está conseguindo identificar muito bem os elementos do conjunto de teste

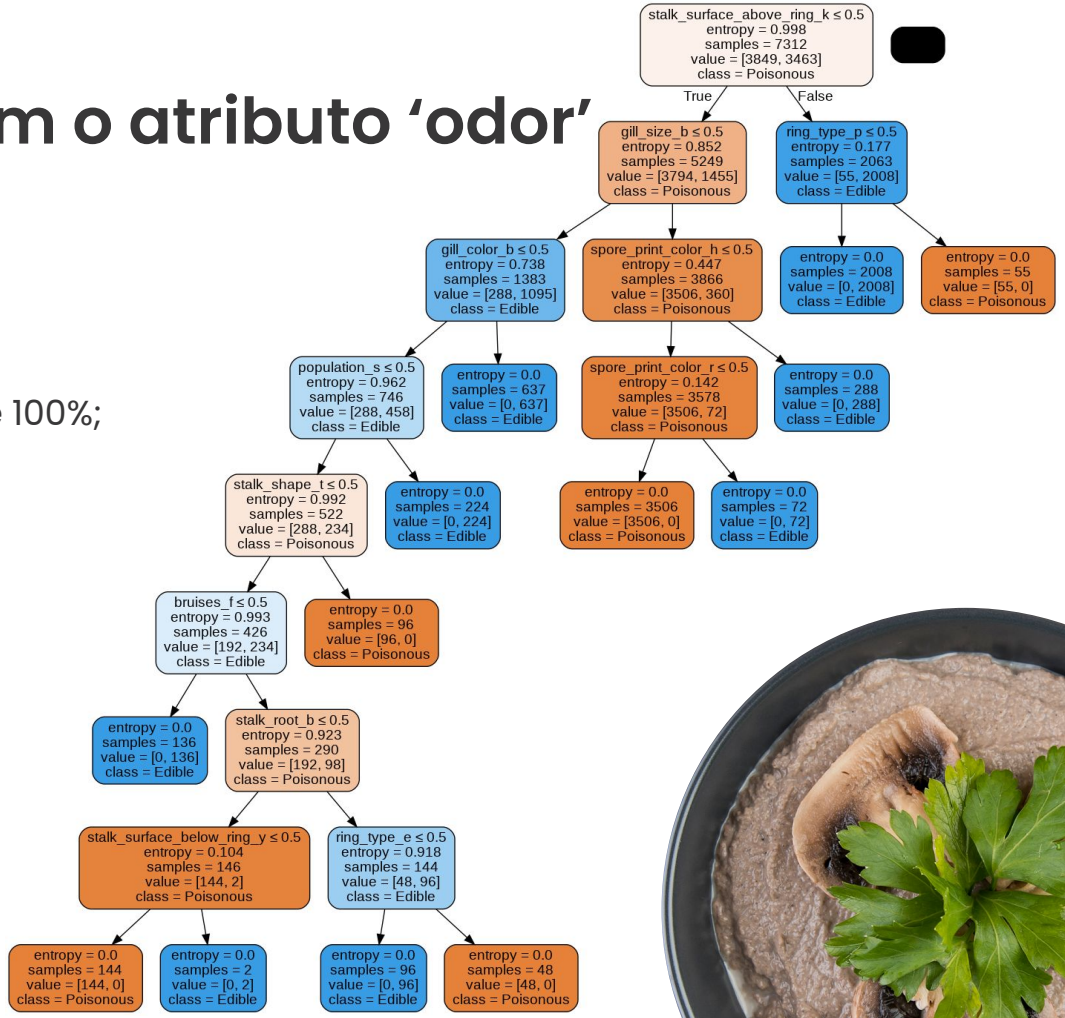
Talvez isso se dê pelo banco de dados ser grande

Outra possibilidade é de que um atributo específico seja responsável por essa classificação tão boa.
No nosso caso, o odor

Repetindo os testes sem o atributo ‘odor’

O k-fold = 10 agora apresentou um valor diferente de 1,
mas continuou com acurácia próxima de 100%;

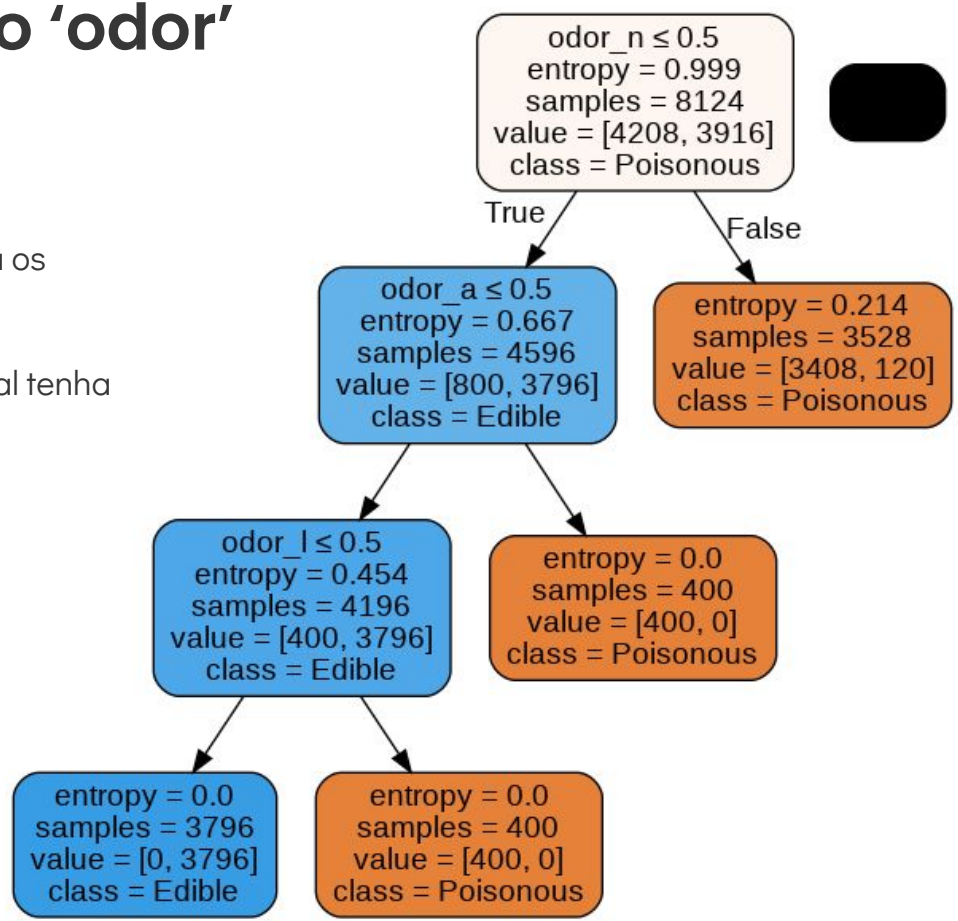
A árvore gerada foi consideravelmente maior.



Criando árvore do atributo ‘odor’

Podemos ver que esse atributo isolado já classifica os cogumelos de uma maneira razoável;

Com isso, é compreensível que a classificação geral tenha sido tão boa.



Variação usando a função de avaliação 'gini'

Novamente, voltamos a utilizar todos os atributos do dataset que não foram eliminados;

Para divisão treinamento/teste, utilizamos k-fold = 5;

A função gini é utilizada para escolher os atributos da árvore.

Com esses parâmetros, obtivemos os seguintes resultados em relação ao conjunto de testes:

Variação usando a função de avaliação 'gini'

Acurácia:

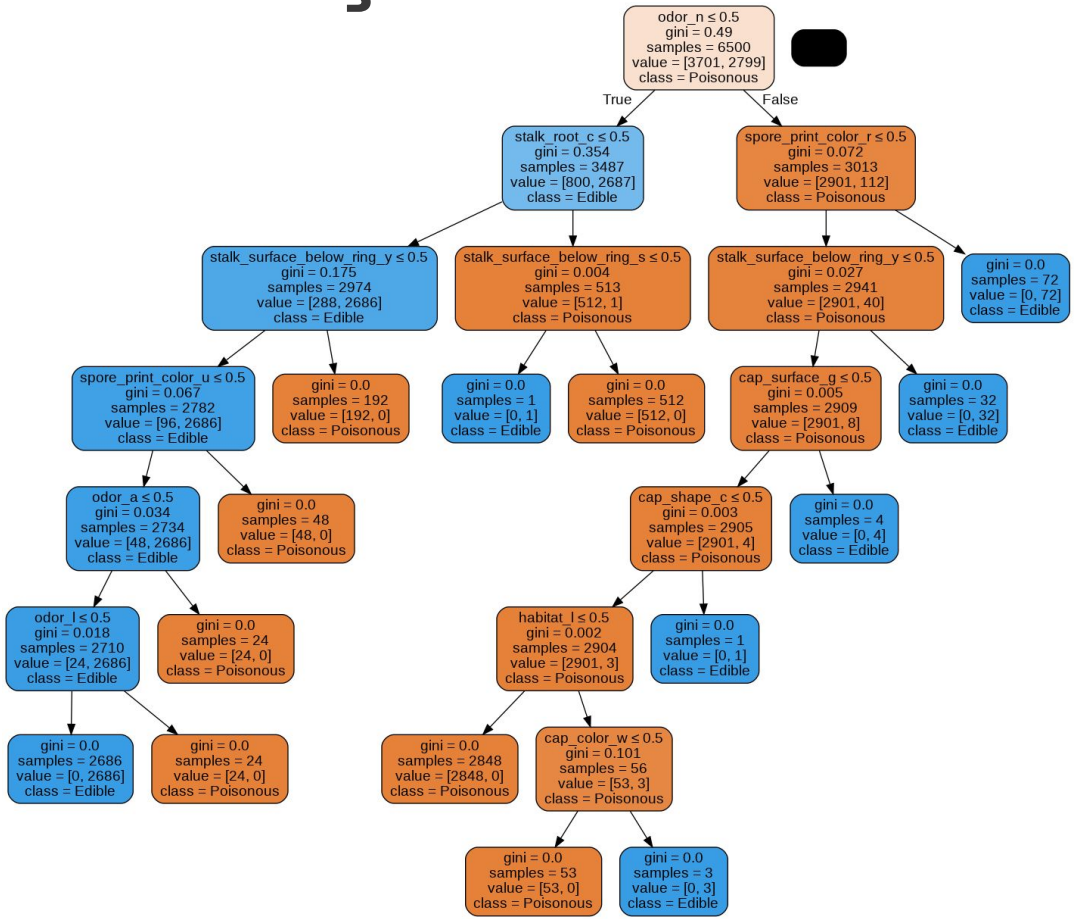
1. 1.0
2. 1.0
3. 1.0
4. 1.0
5. 0.9901477832512315

0.9980295566502463 (média)

Precisão:

1. 1.0
2. 1.0
3. 1.0
4. 1.0
5. 0.9929390997352162

0.9985878199470433 (média)

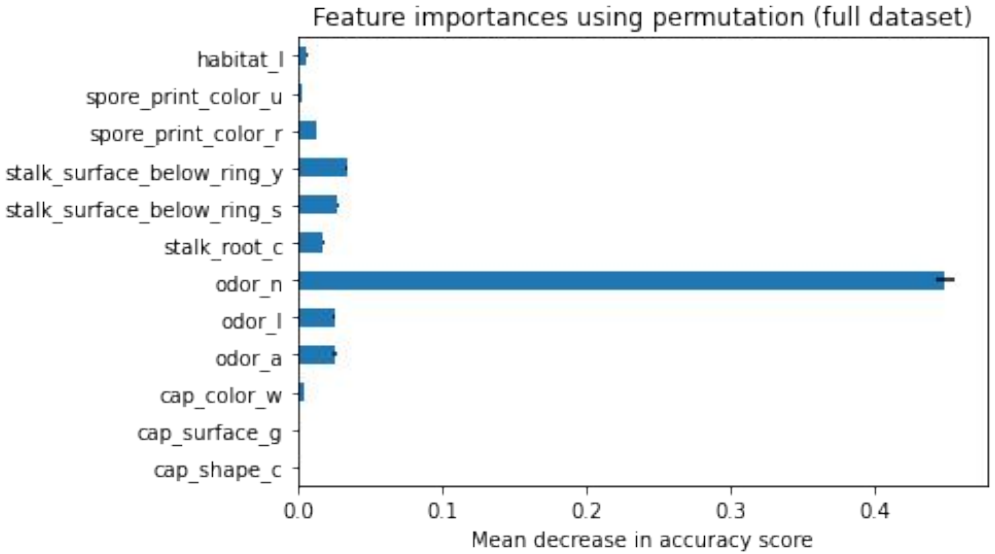


Variação usando a função de avaliação 'gini'

- Matriz com falsos positivos e falsos negativos (test set confusion matrix)

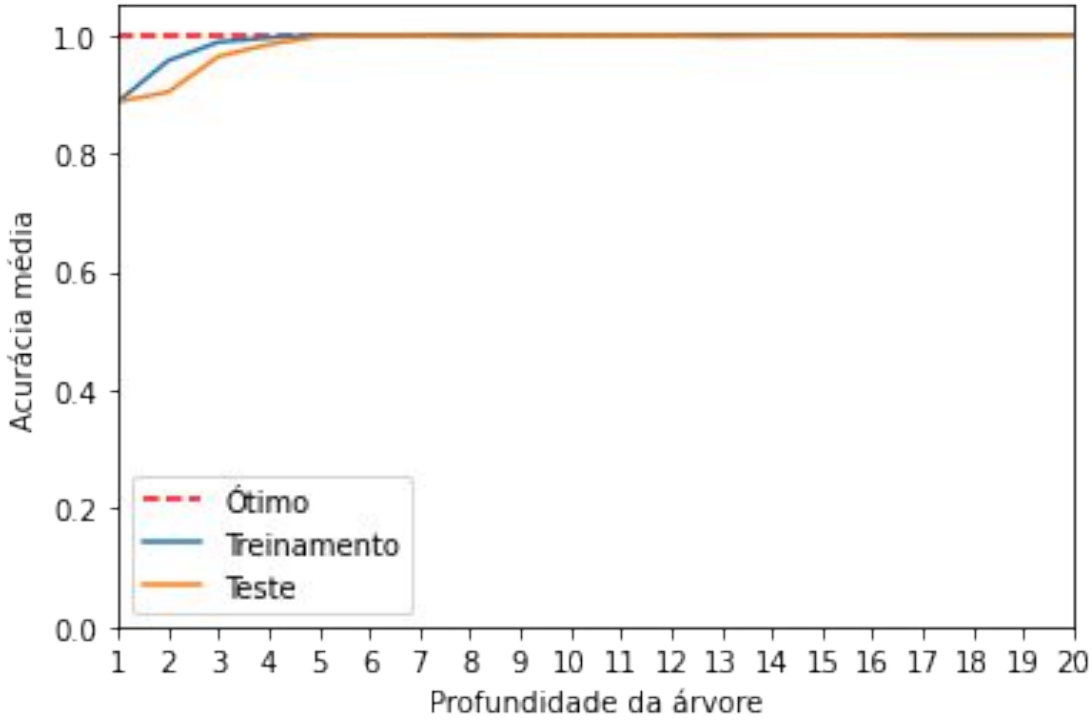
	Predicted edible	Predicted poisonous
Real edible	491	16
Real poisonous	0	1117

- Importância dos atributos



Variação usando a função de avaliação 'gini'

- Podas de profundidade da árvore em relação à acurácia média



Obrigado!