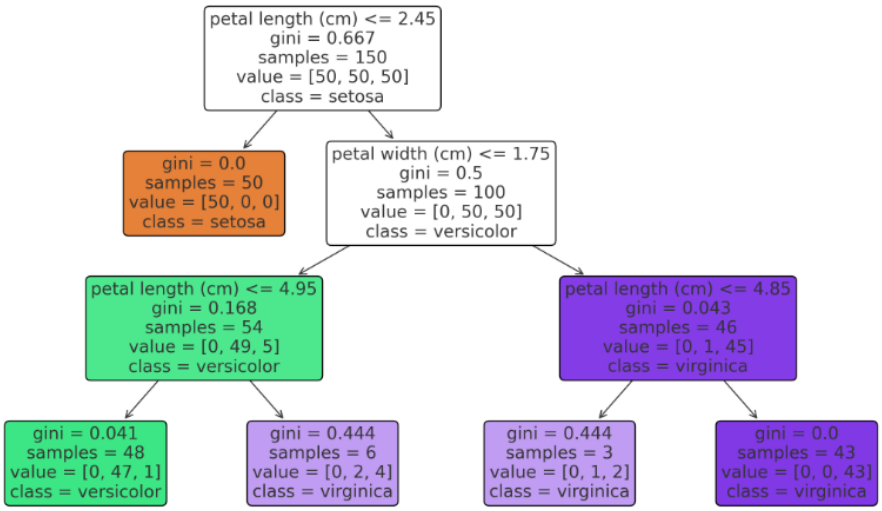
Exercícios de Aprendizado de Máquina I
30-09

Questão 01

A partir da base **Iris** (atributos: *sepal length*, *sepal width*, *petal length*, *petal width*; classes: *setosa*, *versicolor*, *virginica*), foi treinada a árvore de decisão mostrada na **figura** abaixo (gerada com DecisionTreeClassifier, profundidade máxima = 3).



Sobre esta árvore de decisão, avalie:

I. Algumas folhas apresentam **impureza**, indicando mistura de classes.
II. A profundidade 3 **não foi suficiente** para separar todas as instâncias em folhas puras.
III. Os ganhos de informação para os atributos relativos à sépala tiveram valores zero

Quais afirmações estão corretas?

a) Apenas I
b) Apenas I e II
c) Apenas I e III
d) I, II e III
e) Nenhuma é verdadeira

Questão 02

Sobre os métodos de codificação **Label Encoding e One-Hot,** avalie as afirmações abaixo:

I. Em modelos de aprendizado, especialmente aqueles que são baseados em distância, Label Encoding pode induzir ordem e distâncias artificiais entre categorias.
II. One-Hot faz com que qualquer par de categorias distintas tenha distância igual (em Euclidiana), evitando hierarquias artificiais.
III. Para modelos de árvore, One-Hot é sempre superior a Label Encoding.

Quais afirmações estão corretas?

a) Apenas I
b) Apenas I e II
c) Apenas I e III
d) I, II e III
e) Nenhuma é verdadeira

Questão 03

Sobre o funcionamento dos algoritmos ID3, C45 e CART, avalie as afirmações abaixo:

I. O ID3 e o C4.5 podem gerar nós com múltiplas ramificações, dependendo do número de valores de um atributo.
II. O CART gera apenas árvores binárias, independentemente do número de valores de um atributo.
III. O C4.5 transforma atributos contínuos em intervalos binários, de forma semelhante ao CART, mas permite múltiplos ramos para atributos categóricos.

Quais estão corretas?

a) Apenas I
b) Apenas I e II
c) Apenas I e III
d) Apenas II e III
e) I, II e III estão corretas

Questão 04

Considere o seguinte trecho de código:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split

X, y = load_iris(return_X_y=True)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
clf = RandomForestClassifier(n_estimators=100, max_features=2, random_state=0)
clf.fit(X_train, y_train)
```

Com base neste código, analise as seguintes afirmações:

I. O parâmetro max_features=2 limita o número de atributos considerados por divisão em cada árvore para 2.

II. O conjunto de teste reservado é muito alto, o que afetará negativamente o desempenho do modelo

III. O modelo está sujeito a aleatoriedade mesmo com random_state=0, pois o número de árvores é pequeno.

Quais das afirmações estão corretas?

a) Apenas I
b) Apenas II e III
c) Apenas I e III
d) Todas estão corretas
e) Todas estão incorretas

Questão 05

No contexto dos algoritmos de árvore de decisão e Random Forest, avalie as afirmações abaixo:

I. O RandomForestClassifier reduz o risco de overfitting combinando várias árvores de decisão treinadas em subconjuntos diferentes de dados e atributos.
II. No scikit-learn, definir max_depth em uma árvore de decisão ajuda a controlar o overfitting, mas em florestas aleatórias essa configuração é ignorada.
III. Uma única árvore de decisão (DecisionTreeClassifier do scikit-learn) tem alta tendência ao overfitting, especialmente quando não se limita a profundidade.

Quais das afirmações estão corretas?

a) Apenas I
b) Apenas II e III
c) Apenas I e III
d) Todas estão corretas
e) Todas estão incorretas

Questão 06

No contexto dos algoritmos de árvore de decisão e Random Forest, avalie as afirmações abaixo:

I. Em RandomForestClassifier, o parâmetro n_estimators controla o número de árvores treinadas no conjunto.
II. A árvore de decisão do scikit-learn também utiliza o bootstrap aggregating, mesmo quando usada isoladamente.
III. O Random Forest utiliza a técnica de bootstrap aggregating, amostrando dados sem reposição para cada árvore.

Quais das afirmações estão corretas?

a) Apenas I
b) Apenas II e III
c) Apenas I e III
d) Todas estão corretas

e) Todas estão incorretas

Questão 07

Considere o código:

```
from sklearn.metrics import precision_score, recall_score, f1_score

y_true = [1, 0, 1, 1, 0, 1, 0]
y_pred = [1, 0, 1, 0, 0, 1, 1]
```

I. O comando precision_score(y_true, y_pred) retorna a proporção de verdadeiros positivos entre todos os previstos como positivos.
II. O comando f1_score(y_true, y_pred) retorna o mesmo valor que (precision_score + recall_score)/2.
III. O comando recall_score(y_true, y_pred) retorna a proporção de verdadeiros positivos entre todos os positivos reais.

Quais das afirmações estão corretas?

a) Apenas I
b) Apenas I e II
c) Apenas I e III
d) Apenas II e III
e) Todas estão corretas

Questão 08

Considere as seguintes afirmações:

I. Em datasets desbalanceados, a precisão pode ser alta mesmo quando o recall é baixo.
II. O recall é particularmente importante em problemas em que falsos positivos são críticos (ex: diagnóstico de doenças).
III. O F1-score tende a privilegiar o recall em detrimento da precisão.

Quais das afirmações estão corretas?

a) Apenas I
b) Apenas I e II
c) Apenas I e III
d) Apenas II e III
e) Todas estão corretas

Questão 09

Considerando a base de dados abaixo e o algoritmo de **Árvore de decisão ID3**, qual a raiz da árvore e qual o ganho de informação do atributo, respectivamente?

Obs:

1. **É necessário apresentar todos os cálculos. Ou seja, não será considerada questão sem apresentação dos cálculos; 2) A base do log a ser considerada deve ser a quantidade de classes; 3) Preencher os valores das entropias e ganhos de cada atributo na tabela abaixo.**

Atributos de entrada

1) Trabalho Significativo: {Sim, Não}
2) Relações Sociais: {Sim, Não}
3) Equilíbrio Vida-Trabalho: {Sim, Não}

Classe (variável-alvo)

- Nível de Felicidade: {Alto, Médio, Baixo}

ID	Trabalho Significativo	Relações Sociais	Equilíbrio Vida-Trabalho	Nível de Felicidade
1	Sim	Sim	Sim	Alto
2	Sim	Sim	Não	Alto
3	Sim	Sim	Sim	Alto
4	Sim	Sim	Não	Alto
5	Sim	Sim	Não	Alto
6	Sim	Não	Sim	Médio
7	Não	Sim	Sim	Médio
8	Não	Sim	Não	Médio
9	Não	Sim	Sim	Médio
10	Sim	Não	Sim	Médio
11	Sim	Não	Não	Baixo
12	Não	Não	Sim	Baixo
13	Não	Não	Não	Baixo
14	Não	Não	Sim	Baixo
15	Não	Sim	Não	Baixo

a) A raiz da árvore é o atributo **Trabalho Significativo** com ganho de 0,532
b) A raiz da árvore é o atributo **Relações Sociais** com ganho de 0,257
c) A raiz da árvore é o atributo **Equilíbrio Vida-Trabalho** com ganho de 0,069
d) A raiz da árvore é o atributo **Relações Sociais** com ganho de 0,294
e) A raiz da árvore é o atributo **Trabalho Significativo** com ganho de 0,273

Atributo	Entropia	Ganho
Trabalho Significativo		
Relações Sociais		
Equilíbrio Vida-Trabalho		