Questões para fixação - Mineração de dados

1. A mineração de dados pode ser definida como:
   "Extração não trivial de informação implícita, previamente desconhecida, e potencialmente útil".
   (a) Sabe-se que área de mineração é dividida em 4 grandes tarefas, ou seja, problemas que os métodos buscam resolver. Explique quais são as tarefas de mineração de dados, e para cada uma delas, forneça um exemplo de como poderia ser aplicada em algum banco de dados (escreva uma amostra do banco para explicar).
   (b) Como as tarefas estão relacionadas com os conceitos de aprendizado *supervisionado* e *não supervisionado*?
2. Considerando o processo geral para se resolver um problema de aprendizado supervisionado representado na Figura 1, responda o que se pede.

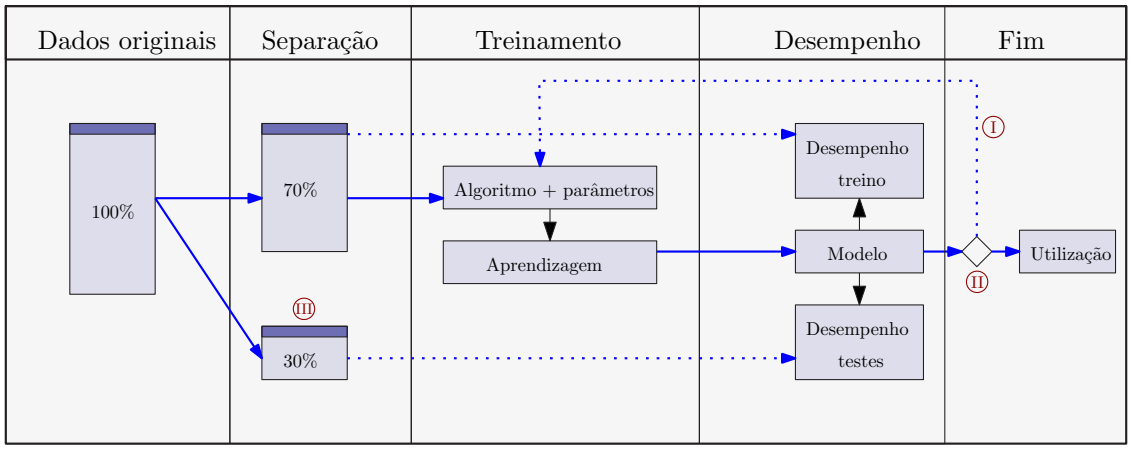


Figura 1: Abordagem geral para resolução de um problema de aprendizado supervisionado

   (a) Explique o qual o que está acontecendo em I, bem como a decisão em II.
   (b) Qual é a importância de se conhecer os parâmetros do algoritmo utilizado para o aprendizado?
   (c) O que está acontecendo em III?
3. O gráfico da Figura 2 mostra a acurácia de um modelo de classificação usando árvores de decisão, para diferentes tamanhos máximos de árvore.
   (a) Pelo gráfico, qual tamanho de árvore você escolheria para o seu modelo?
   (b) O que acontece a partir do de 6 nós?
4. Considere a árvore de decisão representada na Figura 3. A árvore foi treinada para ser um classificador de vinhos (dentre 3 possíveis, *tipo 1*, *tipo 2* e *tipo 3*), com base em 13 características do vinho (representadas em um vetor X).
   Com base nisso, responda o que se pede:
   (a) Qual é o atributo mais importante para diferenciar os vinhos? Por quê?

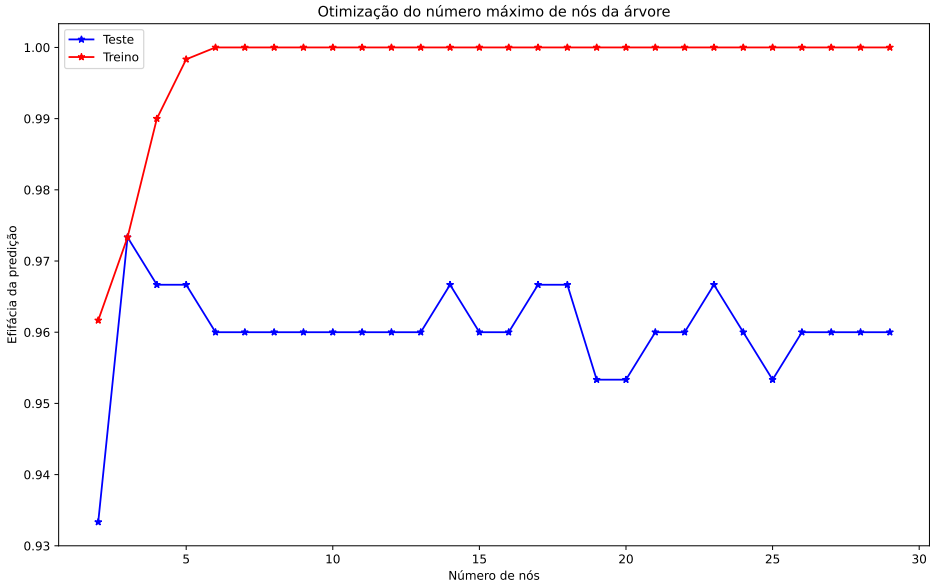


Figura 2: Abordagem geral para resolução de um problema de aprendizado supervisionado

(b) Com base na Figura 3, qual foi o critério usado para a separação dos nós no aprendizado do modelo?

(c) Considerando o vinho com as seguintes características

$$X = [14, 2, 2, 14.56, 128, 2.08, 1, 0.25, 1, 5.5, 1, 10, 700] \tag{1}$$

Como ele seria classificado, de acordo com o modelo? Mostre as condições que você usou na árvore para chegar à classificação final.

(d) Dê um exemplo de vinho (mostrando seu vetor de características X), de forma que ele seja classificado como um vinho do *tipo 1*?

(e) Considere que após a classificação de um conjunto de testes, a seguinte matriz de confusão foi criada.

$$\begin{bmatrix} 14 & 3 & 0 \\ 4 & 22 & 1 \\ 0 & 0 & 10 \end{bmatrix}$$

Em qual classe de vinhos o modelo têm o pior desempenho? Quantas classificações erradas ele fez?

(f) Ainda considerando a matriz de confusão, qual cálculo deveria ser executado para extrair a eficácia do modelo? (não precisa calcular, só deixar o registro da operação e dos elementos envolvidos).

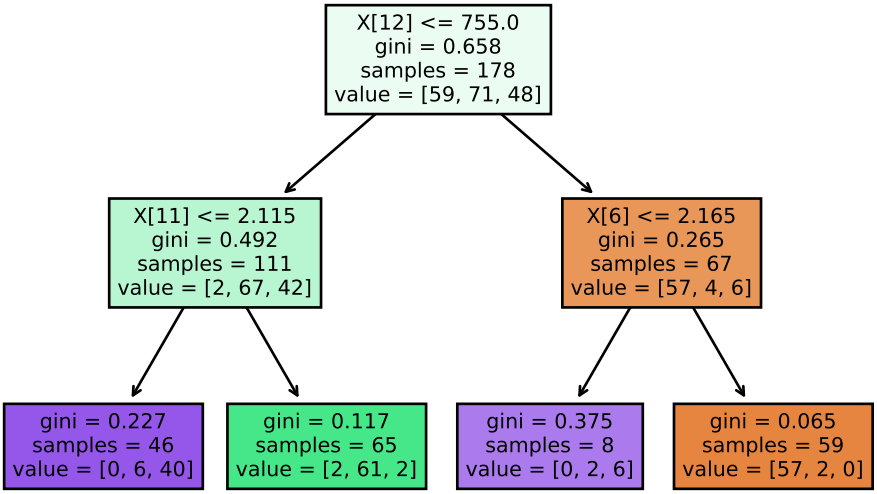


Figura 3: Árvore para classificação de vinhos

5. Sabe-se que a entropia de um nó é dada por:

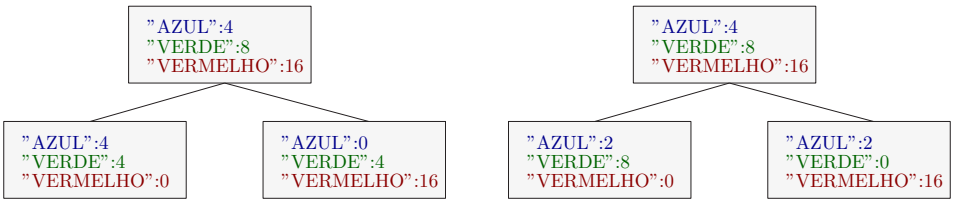
$$E(V) = -\sum_{k} P(v_k) \log_2 P(v_k) \tag{2}$$

Em que k é o número de classes no nó, v_k a k-ésima classe, e $P(v_k)$ a probabilidade de ocorrência da classe v_k no nó. Ainda, o *ganho de informação* ΔI pela separação dos dados de acordo com algum atributo é dado por:

$$\Delta I = E(V_{pai}) - M(E(V_{filhos})) \tag{3}$$

Em que M é a média *ponderada* da entropia dos nós filhos. Considere que dois atributos foram usados para realizar a separação dos dados em um nó, conforme Figura 4. Qual o melhor atributo para separação dos dados, considerando o ganho de informação (responda com base no cálculo)?

6. Considere a amostra de um banco de dados mostrado na Tabela 1. A coluna *Activity* indica a primeira atividade que o paciente foi submetido no hospital, a segunda, terceira e quarta coluna se referem à idade, gênero e peso, respectivamente. O hospital deseja estudar a existência de grupos de pacientes com características semelhantes, de forma a otimizar as atividades executadas, de forma que decidem usar o algoritmo *k-means*.
   (a) Esse algoritmo executa qual tarefa da mineração de dados?



(a) Atributo 1 (b) Atributo 2

Figura 4: 2 atributos usados para separar os dados

(b) Como o banco de dados deve ser transformado para que o algoritmo possa ser aplicado de forma correta?

7. Considere o gráfico gerado na Figura 5. O que este gráfico representa? Qual seria o valor ideal de grupos considerando o gráfico?

Figura 5: Elbow plot

Activity	Age	Gender	Weight
Hospital Admission	56	Female	100
CT Scan	55	Female	50
Discharge	40	Male	57

Tabela 1: Banco de dados atividade inicial em hospital

8. Devido a sua rápida curva de aprendizagem, bem como a uma comunidade ativa e contribuinte, a linguagem de programação Python se tornou muito usada em análise e mineração de dados. Considere às questões abaixo a respeito da linguagem:
   (a) Considerando os dois códigos abaixo, o que será impresso na tela?

```
ed1 = ["A","E","I","O","U"]
ed2 = [("A","E"),("I","O","U"),("F",1,2)]
ed3 = {"A":ed1, "B":ed2 , "C":ed3}
ed4 = ([1,2,3],["a,b,s"], {"C1":[2,22,2], "C2":"C3"})
ed5 = [ed3, ed4]

print(type(ed5))
print(ed5[1][1])
```

```
x = [1,2,3,4,5,6,7]
print(x[5:])
```

   (b) Como você faria para imprimir na tela o elemento "O"contido na estrutura *ed2* a partir da estrutura *ed5*?

9. Para cada um dos códigos abaixo, explique o que ele está fazendo:

```
x = [i for i in range(100) if i%2 == 0]
s = 0
for i in range(len(x)):
        s = s + s[i]
print(s)
```

```
import matplotlib.pyplot as plt
import numpy as np

x1 = np.random.randint(10,20,100)
x2 = np.random.randint(-30,20,100)

fig, ax = plt.subplots(1,2)
ax[0].plot(x1)
ax[1].hist(x2, bins = 5)
plt.show()
```