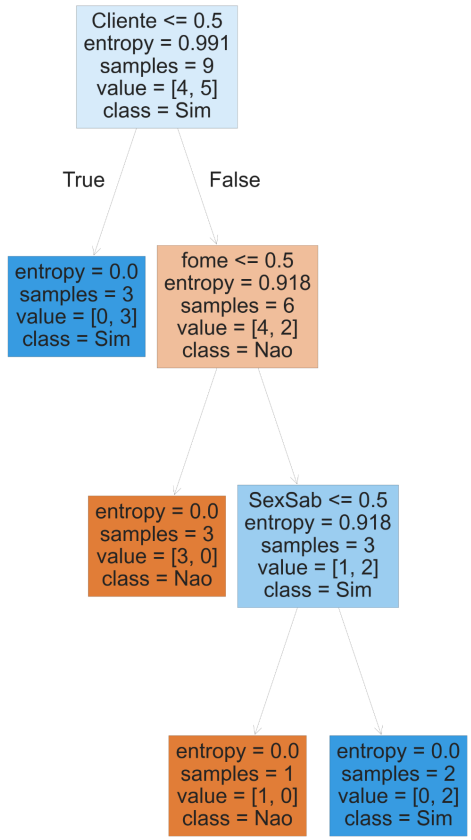
LISTA 3

Questão 1

Código da questão

No código, eu carreguei a base de dados e depois transformei as variáveis categóricas em valores numéricos usando LabelEncoder. Para isso, criei uma lista com as colunas que precisavam ser convertidas, incluindo a coluna Cliente, e apliquei o encoder em todas elas de uma vez. Dessa forma, cada categoria de texto (como “Nenhum”, “Alguns” ou “Cheio”) foi representada por um número inteiro, permitindo que os algoritmos conseguissem interpretar esses dados.

```
#para codificar todos os atributos para laberEncoder de uma única vez
#base_encoded = base.apply(LabelEncoder().fit_transform)
cols_label_encode = ['Alternativo', 'Bar', 'SexSab','fome', 'Cliente', 'Preco', 'Chuva', 'Res','Tempo']
base[cols_label_encode] = base[cols_label_encode].apply(LabelEncoder().fit_transform)
```



Questão 2

[Código da questão]

1. Visualização da base e distribuições

Usei pandas, seaborn e matplotlib para inspecionar os dados. Verifiquei atributos como idade, sexo, classe, tarifa e vi como eles se distribuem em relação à variável alvo (sobrevivência). Isso permitiu identificar quais variáveis eram categóricas e quais eram numéricas.

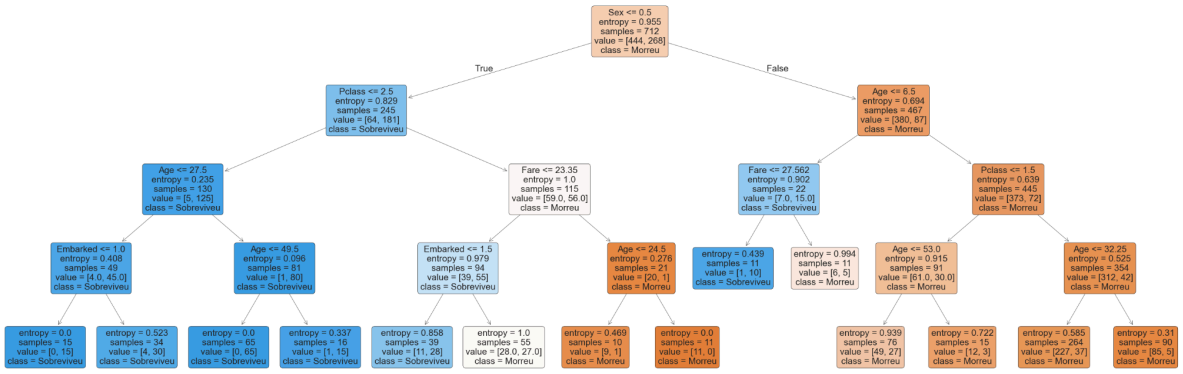
2. Codificação dos atributos

As variáveis categóricas, como sexo e classe, foram transformadas em números usando técnicas de codificação (LabelEncoder). Assim, todos os atributos puderam ser processados pelo algoritmo de árvore de decisão.

3. Regras de mortalidade (padrão encontrado pela árvore)

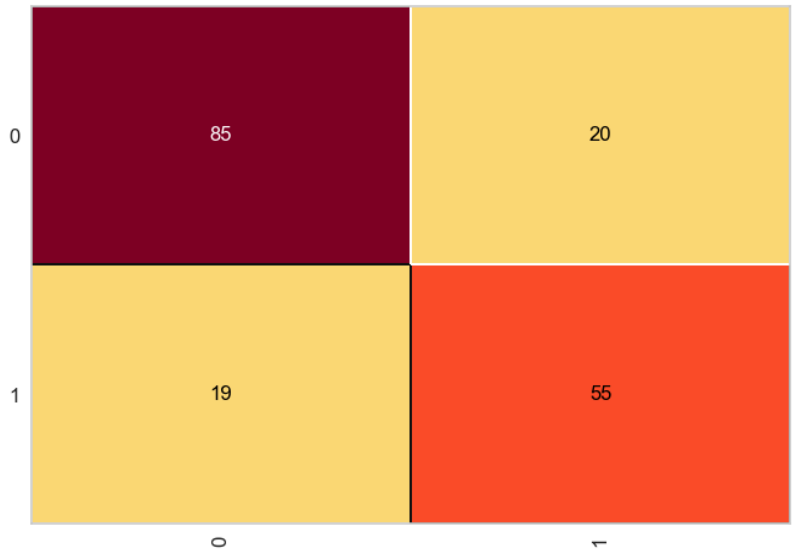
A árvore de decisão gerada mostrou padrões claros, por exemplo:

- Passageiros do sexo feminino tiveram maior taxa de sobrevivência.
- Passageiros da classe 1 também tiveram maiores chances.
- Passageiros do sexo masculino em classes baixas tiveram a maior taxa de mortalidade.



```
print(classification_report(y_teste, previsoes))
```
✓ 0.0s

	precision	recall	f1-score	support
0	0.82	0.81	0.81	105
1	0.73	0.74	0.74	74
accuracy			0.78	179
macro avg	0.78	0.78	0.78	179
weighted avg	0.78	0.78	0.78	179



Questão 3

a) O algoritmo C4.5 é uma extensão do ID3 e foi projetado para superar algumas de suas limitações. As principais diferenças entre os dois são:

- Tipos de dados: O ID3 lida apenas com atributos de dados nominais e não possui suporte para valores ausentes. Já o C4.5 pode usar dados contínuos, lidar com valores desconhecidos e usar atributos com diferentes pesos.
- Sensibilidade a valores: Uma limitação do ID3 é que ele é muito sensível a atributos com um grande número de valores, como um número de CPF, por exemplo. Para superar isso, o C4.5 usa a "taxa de ganho de informação" para medir a proporção do ganho, o que o torna menos propenso a selecionar atributos com muitos valores únicos.
- Pruning (poda): O C4.5 pode podar a árvore de decisão após a sua criação, o que ajuda a reduzir o superajuste dos dados de treinamento e a melhorar o desempenho com novas amostras não vistas. O ID3 não possui essa funcionalidade.
- Desempenho: Em uma comparação de acurácia, o C4.5 teve um desempenho superior ao ID3 para diferentes tamanhos de conjunto de

dados. O C4.5 também demonstrou ser mais rápido que o ID3 em tempo de execução.

b) O algoritmo C4.5 lida com atributos de entrada numéricos (contínuos) da seguinte forma:

1. Ele examina os valores do atributo contínuo nos dados de treinamento.
2. Os valores são ordenados em ordem ascendente.
3. Para cada valor único nesse conjunto ordenado, ele cria uma partição dos registros. Uma partição inclui todos os registros com valores menores ou iguais a um valor de atributo específico, e a outra inclui todos os registros com valores maiores do que esse valor.
4. Para cada uma dessas partições, o C4.5 calcula o ganho ou a taxa de ganho.
5. A partição que maximiza o ganho é então selecionada, e um ponto de corte é definido para o atributo contínuo, transformando-o em um atributo binário.

Essa abordagem permite que o C4.5 use atributos contínuos para a construção da árvore de decisão, algo que o ID3 não faz

Questão 4

LETRA C - Iris_Versicolor, íris_Setosa, Iris_Versicolor, Iris_Virgínica

Questão 5

LETRA C - I e II, apenas

Questão 6

	Precisão	Recall	F1Score	TVP	TFN	TFP	TVN
A	10/17	10/17	0,5882	10/17	7/17	7/105	98/105
B	15/23	15/18	0,7317	15/18	3/18	8/106	98/106
C	20/26	20/30	0,7143	20/30	10/30	6/92	86/92
D	50/56	50/57	0,885	50/57	7/57	6/65	59/65