Lista 03
Disciplina: Inteligência Artificial
Profa.: Cristiane Neri Nobre
Data: 29/08/2025
Aluna: Alessandra Faria Rodrigues
Matrícula: 828333

Questão 01

Considerando-se os códigos (Lendo_e_tratando_arquivo_v2.ipynb e DecisionTree_Restaurante.ipynb.ipynb) disponibilizados no CANVAS, pede-se:

1) Gerar a árvore para a base de dados Restaurante, alterando a codificação do atributo cliente para conter a seguinte codificação:
Cliente_Nenhum = 0
Cliente_Algum = 1
Cliente_Cheio = 2

Código disponível em:
https://github.com/ale-faria/CC/tree/main/IA/Lista03/Python%20Restaurante

Breve discussão:

1. Resumo do Processo

O objetivo foi criar um modelo de Machine Learning para prever se um cliente deve esperar ou não em um restaurante com base em um conjunto de atributos. O processo incluiu:

- **Pré-processamento de Dados**: Tratamento de variáveis com diferentes técnicas (mapeamento manual para Cliente, LabelEncoder e OneHotEncoder para os demais).
- **Divisão dos Dados**: Separação da base em 80% para treino e 20% para teste (*Holdout*).
- **Treinamento e Avaliação**: Utilização do algoritmo de Árvore de Decisão (critério='entropy') e avaliação de sua performance no conjunto de testes.

2. Análise dos Resultados

- O modelo atingiu uma acurácia geral de 66.7% no conjunto de teste.
- Verdadeiros Negativos (Acertou "Não"): 1
- Falsos Positivos (Errou, prevendo "Sim" quando era "Não"): 1
- Falsos Negativos (Errou, prevendo "Não" quando era "Sim"): 0
- Verdadeiros Positivos (Acertou "Sim"): 1
- Recall da classe “Sim” foi de 100% e da classe “Não” foi de 50%

- Precisão da classe “Sim” foi de 50% e da classe “Não” foi de 100%

3. Conclusões: O conjunto de teste tinha apenas 3 amostras e a base de dados era muito pequena, o que torna os resultados não muito relevantes. O modelo provavelmente "decorou" os dados de treino, criando regras muito específicas para cada exemplo, o que prejudicou sua capacidade de generalizar para os novos dados do conjunto de testes.

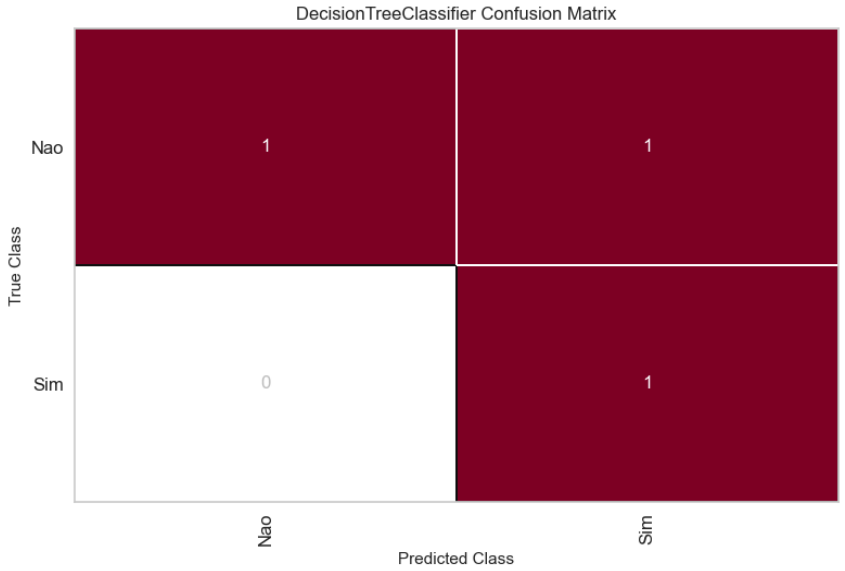

Figura 1: matriz de confusão

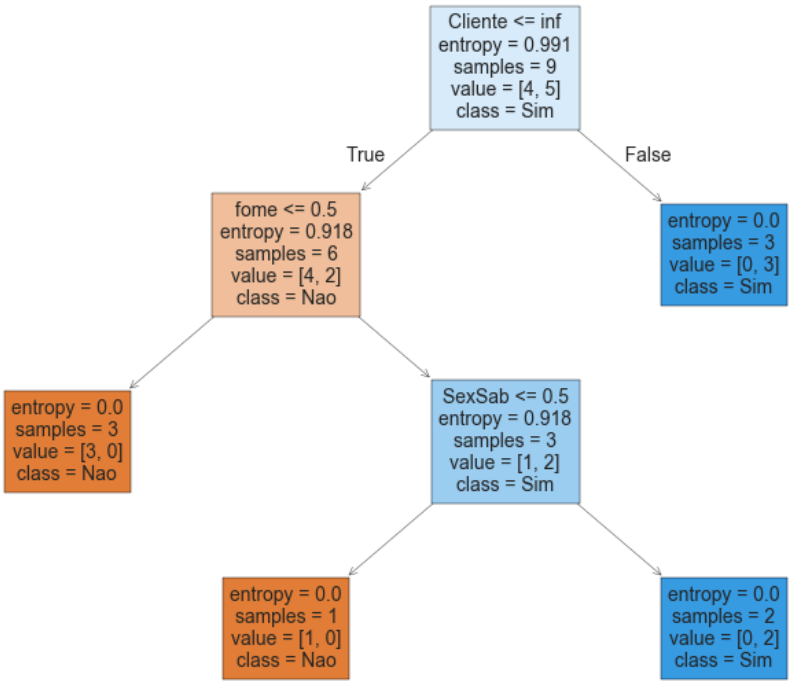

Figura 2: árvore de decisão gerada a partir da nova codificação

Questão 02

Baseado nestes códigos acima, encontrar o padrão de pessoas que sobreviveram ao desastre do TITANIC, que matou mais de 1.500 pessoas em 1912. A base de dados do TITANIC está no CANVAS.

1. Visualize a base de dados primeiro, veja como estão os atributos e suas distribuições.
2. Investigue a melhor forma de codificar cada atributo da base de dados.
3. Forneça as regras que mostre o padrão de mortalidade.

Breve discussão:
Código de visualização dos dados disponível em:
https://github.com/ale-faria/CC/blob/main/IA/Lista03/Titanic/visualizacao.py
Gráficos gerados disponível em:
https://github.com/ale-faria/CC/tree/main/IA/Lista03/Titanic/Graficos

Ao analisar os gráficos, foi possível obter algumas informações importantes sobre os dados:

- **Survived**: Mais pessoas não sobreviveram (0) do que sobreviveram (1).
- **Pclass (Classe)**: A grande maioria dos passageiros viajava na 3ª classe, seguida pela 1ª e pela 2ª.
- **Sex (Sexo)**: Havia bem mais passageiros do sexo masculino do que do feminino.
- **SibSp (Irmãos/Cônjuges a Bordo)**: A maioria das pessoas viajava sem irmãos ou cônjuges.
- **Parch (Pais/Filhos a Bordo)**: A maioria viajava sem pais ou filhos.
- **Age (Idade)**: O histograma mostra uma concentração de passageiros jovens, especialmente entre 20 e 30 anos, com poucas pessoas acima de 60 anos. uma porção significativa dos dados de idade está faltando e será tratada em uma análise posterior.
- **Cabin (Cabine)**: a maioria dos valores para a cabine está ausente.

Código de pré processamento de dados disponível em:
https://github.com/ale-faria/CC/blob/main/IA/Lista03/Titanic/processamento.py
Código da árvore de decisão disponível em:
https://github.com/ale-faria/CC/blob/main/IA/Lista03/Titanic/arvore.py

Resultados do Modelo de Árvore de Decisão:

- **Acurácia**: O modelo acertou **79.89%** das previsões no conjunto de dados de validação.
- **Não Sobreviveu**: O modelo é bom em identificar quem não sobreviveu. Ele acertou 94% dos casos reais (recall) e, quando previu que alguém não sobreviveria, estava certo 78% das vezes (precision).

- **Sobreviveu**: O desempenho é um pouco mais baixo para prever quem sobreviveu. Ele conseguiu identificar corretamente 58% dos sobreviventes reais (recall). No entanto, quando o modelo previu que alguém sobreviveria, ele estava certo 85% das vezes (precision).
- **Matriz de Confusão**:
    - 103 (Verdadeiro Negativo): Passageiros que não sobreviveram, e o modelo previu corretamente.
    - 40 (Verdadeiro Positivo): Passageiros que sobreviveram, e o modelo previu corretamente.
    - 7 (Falso Positivo): O modelo previu que sobreviveriam, mas eles não sobreviveram.
    - 29 (Falso Negativo): O modelo previu que não sobreviveriam, mas eles sobreviveram.

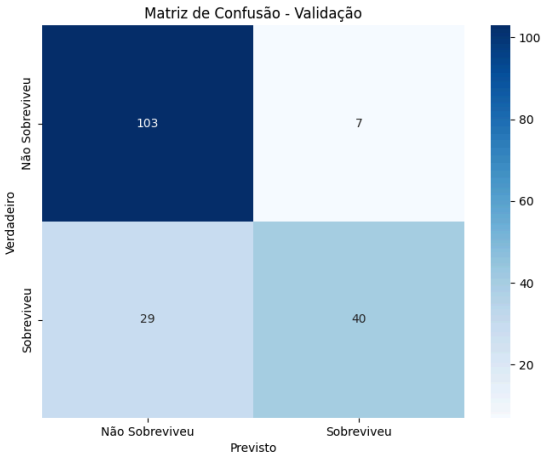

Figura 3: matriz de confusão

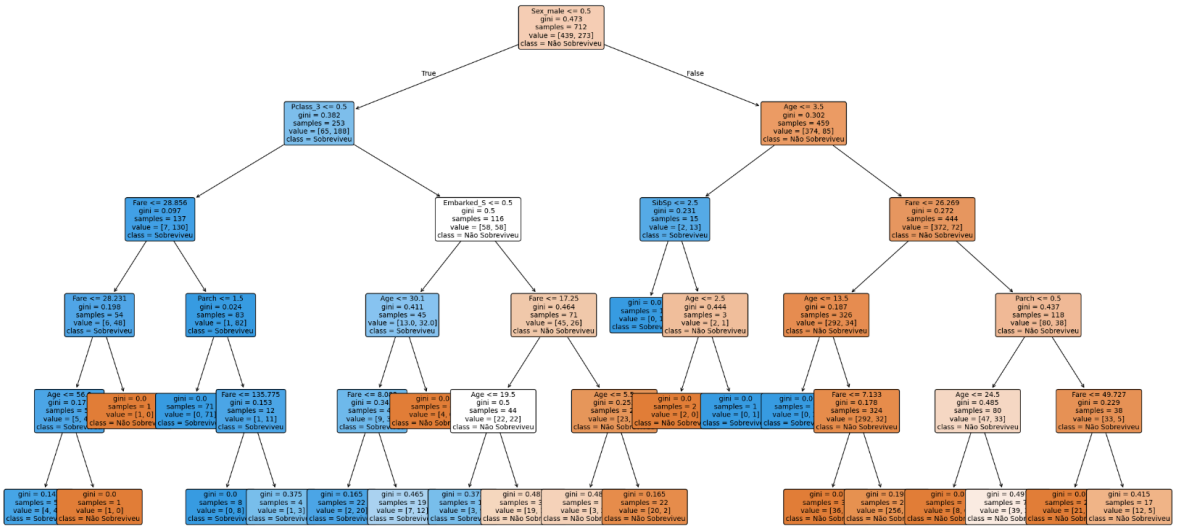

Figura 4: árvore de decisão gerada

Questão 03

Leia o artigo “A_comparative_study_of_decision_tree_ID3_and_C4.5.pdf” que está no CANVAS e responda:

1) Quais as diferenças entre os algoritmos de árvore ID3 e C4.5?

O algoritmo C4.5 é uma evolução do ID3 e as principais diferenças são:

i) **Tratamento de Dados Contínuos:** O C4.5 pode lidar com atributos de valores contínuos (numéricos). Ele faz isso ordenando os valores e encontrando o melhor ponto de divisão para maximizar o ganho de informação. O ID3, por outro lado, geralmente utiliza apenas atributos nominais.

ii) O C4.5 possui um mecanismo para tratar instâncias de **dados com valores ausentes** ou desconhecidos. O ID3 não lida com valores ausentes.

iii) O C4.5 implementa a **poda da árvore** após a sua criação.

iv) O ID3 usa o **Ganho de Informação** para selecionar o atributo de divisão, o que pode favorecer indevidamente atributos com um grande número de valores distintos. Para resolver esse problema, o C4.5 utiliza a **Razão de Ganho** (Gain Ratio), que normaliza o Ganho de Informação e resulta em uma seleção de atributos mais justa.

v) O C4.5 tem a capacidade de utilizar atributos com **pesos** diferentes, permitindo que alguns atributos sejam considerados mais importantes que outros no processo de classificação.

2) Como o algoritmo C4.5 lida com os atributos de entrada que são numéricos?

O C4.5 lida com atributos numéricos escolhendo pontos de corte (thresholds) que transformam o atributo contínuo em um teste binário, e seleciona o corte que gera a melhor separação entre as classes.
O algoritmo avalia todos os possíveis pontos de corte entre valores diferentes do atributo. Para cada candidato, calcula-se a Razão de Ganho (gain ratio), que mede a pureza das classes após a divisão e o ponto que maximiza o ganho é escolhido como o corte.

Questão 04

Analisando a tabela e fazendo o caminho pela árvore gerada, a resposta correta é letra c
c) Iris_Versicolor, íris_Setosa, Iris_Versicolor, Iris_Virgínica

Questão 05

Considerando a árvore da questão anterior, e as seguintes afirmações:

I. Esta árvore possui 5 regras de classificação
II. Das regras geradas, há apenas uma com cobertura por classe de 100%
III. A menor cobertura por classe é de 6.8% e corresponde à classe Iris_Virgínica

É correto o que se afirma em:
a) I, apenas.
b) III, apenas.
c) I e II, apenas.
d) I e III, apenas.
e) I, II e III.

I está correta, já que a árvore possui 5 nós folhas.
II está correta, somente a classificação das serosas tem uma cobertura de 100%
III está incorreta, a menor cobertura é de 2,7% da iris_versicolor

Questão 06

Considere a seguinte matriz de confusão obtida por meio do classificador, Árvore de decisão, para um problema de quatro classes:

		Foi classificado como			
		A	**B**	**C**	**D**
Era da classe	**A**	10	4	2	1
	B	1	15	2	0
	C	2	3	20	5
	D	4	1	2	50

Quais os valores para as métricas abaixo para cada uma das classes A, B, C e D?

	Precisão	**Recall**	**F1Score**	**TVP**	**TFN**	**TFP**	**TVN**
A	**10/17~0,6**	**10/17~0,6**	**0,6**	**10/17**	**7/17**	**7/105**	**98/105**
B	**15/23~0,7**	**15/18~0,8**	**0,74**	**15/23**	**8/23**	**8/104**	**96/104**
C	**20/26~0,8**	**20/30~0,7**	**0,74**	**20/26**	**6/26**	**6/92**	**86/92**
D	**50/56~0,9**	**50/57~0,9**	**0,9**	**50/56**	**6/56**	**6/65**	**59/65**