Lista #4

Curso: Ciência da Computação

Disciplina: Inteligência Artificial

Profª. Cristiane Neri Nobre

Data de entrega: 24/09

Valor: 3 pontos

Objetivo

Implementar, comparar e justificar as escolhas de projeto de três algoritmos de árvore de decisão:

- ID3 (ganho de informação; atributos categóricos),
- C4.5 (razão de ganho; lida com contínuos),
- CART (índice de Gini; divisões binárias).

Você deverá:

1. implementar os três algoritmos do zero (sem usar DecisionTreeClassifier para treinar; ele pode ser usado apenas como baseline de verificação);
2. explicar todas as decisões técnicas tomadas;
3. mostrar as saídas (árvore aprendida, métricas e artefatos pedidos) para os dataset descritos abaixo

Observação:

Para todas as listas que envolvem a implementação das funções em Python, solicita-se que:

1. Todas as discussões das questões devem estar contidas na lista. Ou seja, todas as decisões e explicações necessárias precisam estar na lista, que só pode ser entregue em PDF. Listas em qualquer outro formato serão zeradas.
2. Os links para os códigos desenvolvidos devem estar inseridos na lista, com as devidas permissão de acesso. Listas sem permissão de acesso serão zeradas.
3. Este código desta lista deverá estar disponível como biblioteca do python. Mostre como instalar a biblioteca e mostre também todas as saídas dos códigos no próprio PDF.
4. Adicione o link da biblioteca no PDF

Questão 01

Dataset 1 sugerido

Titanic — sobrevivência (Kaggle).

- Target: Survived (0/1)
- Atributos (use ao menos estes): Pclass (1/2/3), Sex, Age, SibSp, Parch, Fare, Embarked (C/Q/S).

- Motivos: mistura de variáveis categóricas e contínuas; classe desbalanceada moderada; fácil interpretação.

Dataset 2 sugerido

Play Tennis (14 linhas, atributos categóricos), clássico para conferir se o ID3/C4.5 estão corretos antes de ir ao Titanic.
Atributos: Outlook (Sunny/Overcast/Rain), Temperature (Hot/Mild/Cool), Humidity (High/Normal), Wind (Weak/Strong).
Classe: Play (Yes/No).

Entregáveis (um único notebook/relatório)

1. Seção 1 — Preparação dos dados
    - Titanic: limpeza de missing values (Age, Embarked, etc.).
    - Partição: train/ test (ex.: 80/20, estratificada).
    - ID3: justificar discretização de contínuos (p.ex., Age, Fare) —discretization simples.
    - C4.5/CART: contínuos tratados nativamente por limiares.
2. Seção 2 — Implementações

2.1 Utilidades comuns

- Cálculo de entropia, ganho de informação, razão de ganho, Gini.
- Procura de melhor divisão:
    - Categóricos: por valor (ID3/C4.5) ou binarização ótima (CART).
    - Contínuos: varredura por limiar (ordenar valores únicos; testar pontos médios entre adjacentes).
- Empates (mesmo critério): explicar os critérios de empate

2.2 ID3 (do zero)

- Critério: ganho de informação.
- Atributos categóricos: use o Titanic discretizado.

2.3 C4.5 (do zero)

- Critério: razão de ganho (ganho normalizado pela entropia do split).
- Contínuos: selecionar limiar ótimo; nós multi-ramificados para categóricos.
- Handling de missing: média e moda

2.4 CART (do zero)

- Critério: Gini; splits binários sempre.
- Compare sua árvore com sklearn.tree.DecisionTreeClassifier(criterion="gini").

3. **Seção 3 — O que mostrar como "saídas"**

Para cada algoritmo (ID3, C4.5, CART), inclua a árvore gerada e mostre as regras obtidas