**Questão 1 - Gini vs Entropy:** O critério Gini mede a impureza de um nó (a probabilidade de classificar erroneamente um elemento). Se todas as amostras pertencem à mesma classe, Gini = 0 (nó puro). Para um nó com proporção p de classe positiva e (1-p) de negativa:

Gini = 1 - (p² + (1-p)²)

Por exemplo, se um nó tem 7 exemplos da classe 0 e 3 da classe 1:

* p(classe 0) = 0.7
* p(classe 1) = 0.3
* Gini = 1 - (0.7² + 0.3²) = 1 - (0.49 + 0.09) = 1 - 0.58 = 0.42

O Entropy (usado no seu modelo final) mede a aleatoriedade ou desordem: Entropy = -p₁log₂(p₁) - p₂log₂(p₂)

A principal diferença entre ambos: Entropy é computacionalmente mais custoso devido ao cálculo de logaritmos, mas muitas vezes pode resultar em árvores mais balanceadas.

**Questão 2 - Hiperparâmetros:**

* [Tratamento dos Dados](https://github.com/joaodest/inteligencia-artificial/blob/master/Lista%202/Lendo_e_tratando_arquivo_v2.ipynb)
* [Modelo](https://github.com/joaodest/inteligencia-artificial/blob/master/Lista%202/DecisionTree_com_gridsearch.ipynb)
* **max\_depth=3**: Limita a profundidade para evitar overfitting. Na nova implementação da árvore, 3 níveis foram suficientes para capturar os padrões importantes.
* **max\_features=0.8**: Considera 80% das features em cada divisão, introduzindo alguma aleatoriedade e reduzindo correlações.
* **min\_samples\_leaf=6**: Exige que cada nó folha tenha pelo menos 6 amostras, evitando decisões baseadas em poucos exemplos.
* **min\_samples\_split=2**: Valor mínimo que permite divisões fáceis (2 é o padrão).

**Questão 3 - Otimizadores:**

* **GridSearchCV**: Testa todas as combinações de parâmetros especificadas (abordagem exaustiva). Foi o que usamos.
* **RandomizedSearchCV**: Testa aleatoriamente um subconjunto de combinações, útil quando o espaço de parâmetros é grande.
* **BayesSearchCV**: Usa processo gaussiano para direcionar a busca, aprendendo com iterações anteriores quais áreas do espaço de parâmetros são mais promissoras.

**Questão 4 - Métricas em classificação de fraudes:** A resposta é A (I e II são corretas).

* I: Alta precisão significa poucos falsos positivos, mas pode haver muitos falsos negativos.
* II: Alta revocação significa poucos falsos negativos, mas pode haver muitos falsos positivos.
* III: F1 é a média harmônica (não aritmética) de precisão e revocação.

**Questão 5 - Métricas em diagnóstico:** A resposta é A (I e II são corretas).

* I e II são corretas pelas mesmas razões da questão anterior.
* III é falsa porque para doenças letais, alta revocação (não perder nenhum caso positivo) geralmente é mais importante que precisão.

**Questão 6 - C4.5 vs ID3:** Principais diferenças:

* C4.5 usa razão de ganho (gain ratio) enquanto ID3 usa ganho de informação
* C4.5 lida com atributos contínuos, ID3 não
* C4.5 lida com valores ausentes
* C4.5 inclui poda de árvore
* C4.5 pode lidar com diferentes pesos para atributos

**Questão 7 - Razão de Ganho vs Ganho de Informação:**

* Ganho de Informação: IG(S,A) = Entropy(S) - Σ(|Sv|/|S|) \* Entropy(Sv)
* Razão de Ganho: GR(S,A) = IG(S,A) / SplitInfo(S,A)

Onde SplitInfo(S,A) = -Σ(|Sv|/|S|) \* log₂(|Sv|/|S|)

A razão de ganho normaliza o ganho de informação pela entropia da divisão do atributo, corrigindo o viés do ganho de informação para atributos com muitos valores possíveis.