Machine Learning com Python

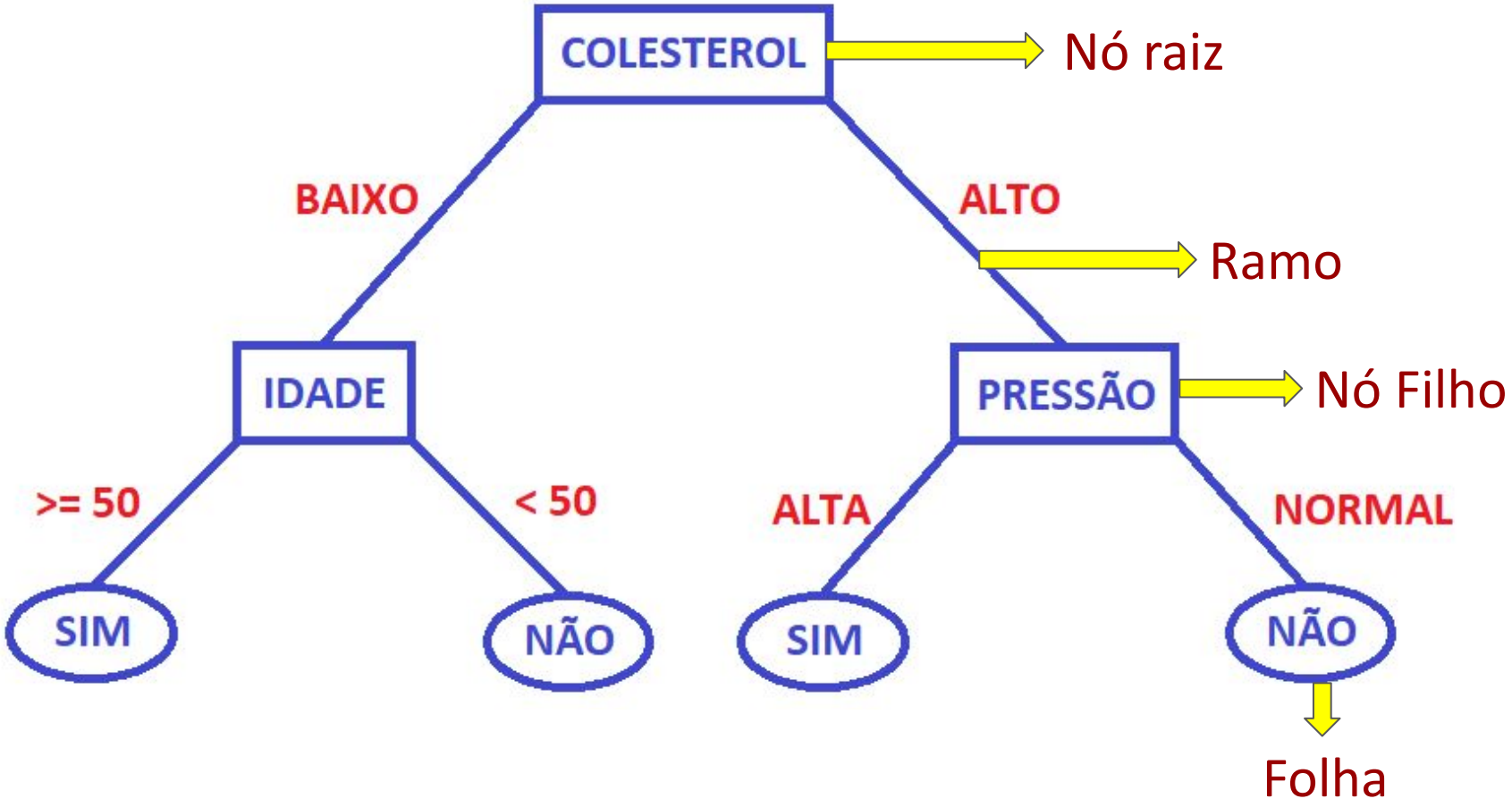
Prof. Luciano Galdino

ÁRVORES DE DECISÃO (Decision Tree)

Aplicado em problemas de aprendizagem supervisionada tanto de classificação (mais utilizado) como de regressão.

Seleciona a ordem que os atributos irão aparecer na árvore, sempre de cima para baixo, conforme sua importância para a predição, assim como determina a separação dos ramos da árvore.

Estrutura similar a um fluxograma. É composto por nós, ramos e folhas.


COLESTEROL
Nó raiz
BAIXO
ALTO
Ramo
IDADE
PRESSÃO
Nó Filho
>= 50
< 50
ALTA
NORMAL
SIM
NÃO
SIM
NÃO
Folha

Para determinar o nível de importância de um atributo, denominado de **ganho de informação**, utiliza-se de várias métricas, sendo que as mais aplicadas são a **entropia** (medida da falta de homogeneidade) e o índice de Gini (medida do grau de heterogeneidade).

Cálculo da Entropia (E):

$$E(S) = -\sum_{i=1}^{n} p_i \cdot log_2(p_i)$$

Cálculo do Índice de Gini:

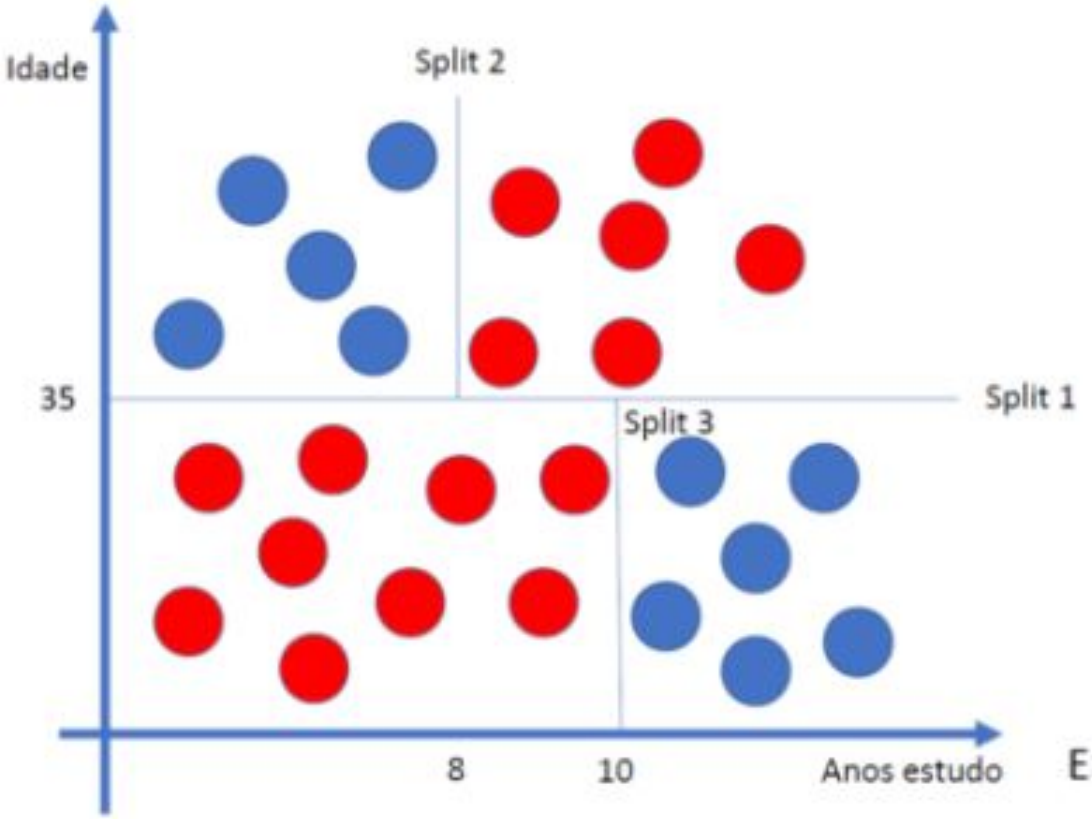
$$\text{Índice Gini} = 1 - \sum_{i=1}^{n} p_i^2$$

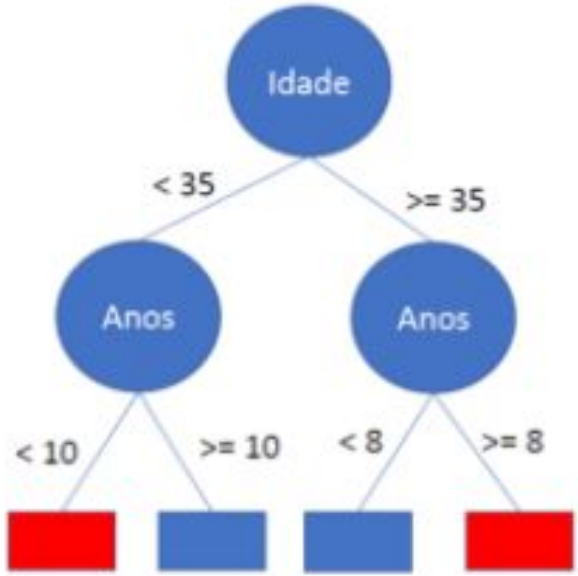
p_i = probabilidade de ocorrência do atributo nos dados.
n = número de classes que pode ser atingida.

Cálculo do Ganho de informação (G):

$$G(S,A) = E(S) - \sum_{i=1}^{n} p(a) \cdot P(i|a) \cdot log_2(P(i|a))$$

p(a) = probabilidade de ocorrência de um categoria de um determinado atributo.
p(i/a) = probabilidade da classe i ocorrer, dado que a já tenha ocorrido.
n = número de classes que pode ser atingida.





Encontrar o melhor conjunto de divisores

https://medium.com/@msremigio/%C3%A1rvores-de-decis%C3%A3o-decision-trees-4cb6857671b3

PODAGEM DAS ÁRVORES

Objetiva diminuir a probabilidade de overfitting.

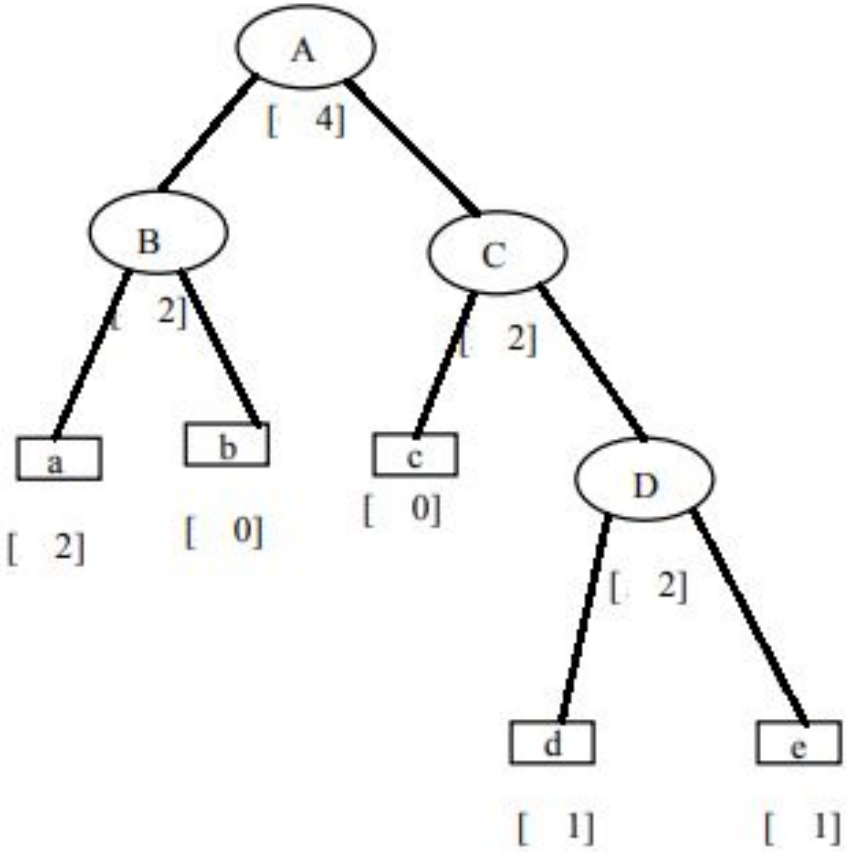
Pode ser de duas formas:

1) Pré-podagem: parar o crescimento da árvore.
2) Pós-podagem: poda com a árvore já completa.

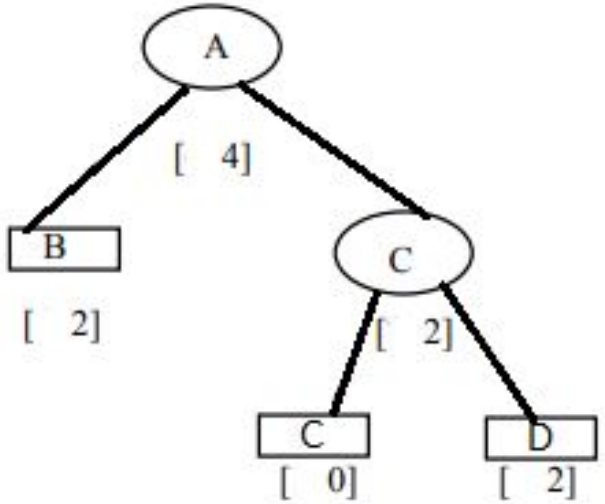
Processo de podagem:

- Percorre a árvore em profundidade.
- Para cada nó de decisão calcula o erro no nó e a soma dos erros nos nós descendentes.
- Se o erro do nó é menor ou igual à soma dos erros dos nós descendentes então o nó é transformado em folha.

Árvore original



Árvore após poda



https://www.maxwell.vrac.puc-rio.br/7587/7587_4.PDF

Vantagens

- Fácil entendimento e interpretação.
- Normalmente não necessitam de preparações sofisticadas nos dados (label Encoder e OneHot Encoder).
- Trabalha com valores faltantes, variáveis categóricas e numéricas.
- Atua com dados não linearmente separáveis.

Desvantagens

- Sujeito a problemas de overfitting.
- Os modelos são instáveis (possuem alta variância).
- Não garante a construção da melhor estrutura para os dados de treino em questão (Necessita treinar várias árvores distintas).