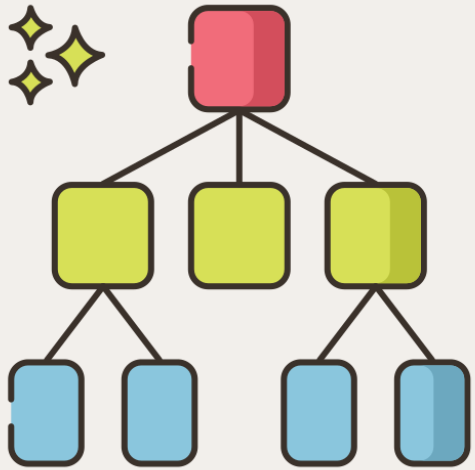
ANÁLISE INTELIGENTE DE DADOS (COB 754)

ÁRVORES DE DECISÃO

LETÍCIA MARTINS RAPOSO

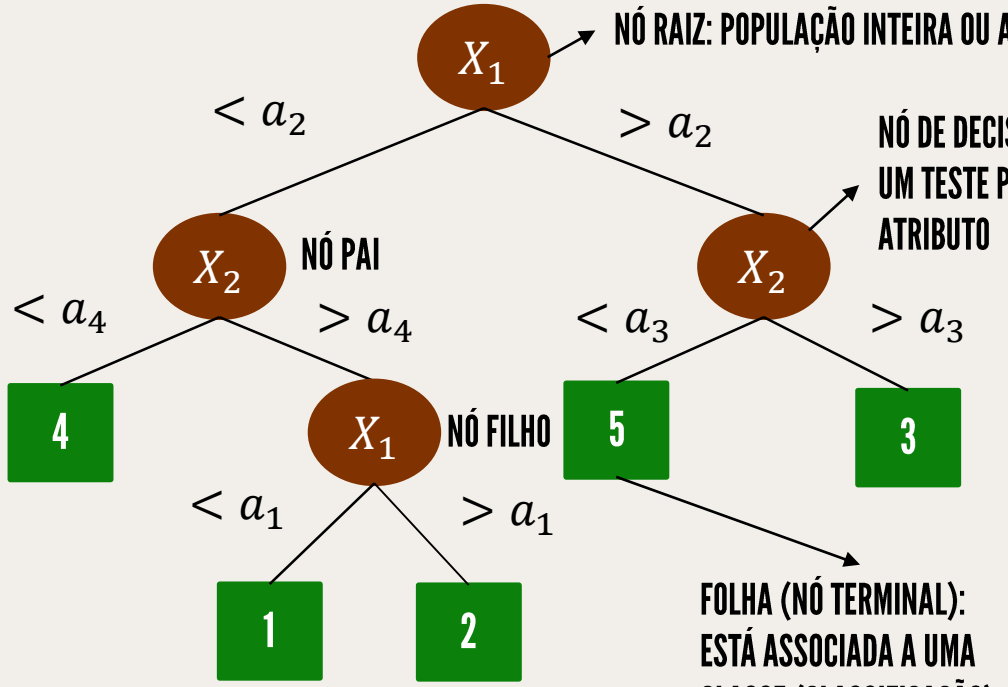
ÁRVORES DE DECISÃO

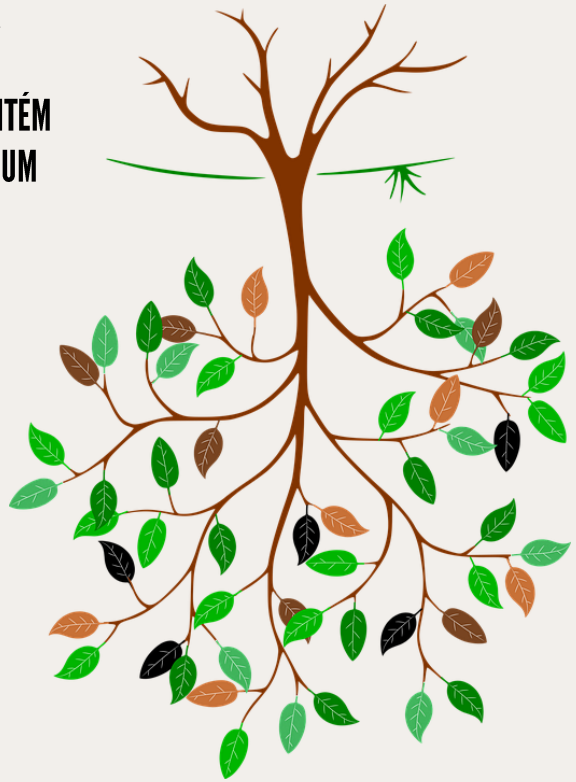


CARACTERÍSTICAS

- SUPERVISIONADO
- USADO PRINCIPALMENTE EM CLASSIFICAÇÃO
- DIVIDIR PARA CONQUISTAR
- REGRAS DO TIPO SE-ENTÃO
- DIVISÃO DO ESPAÇO EM SUBESPAÇOS
- ABORDAGEM "DE CIMA PARA BAIXO"
- "GANANCIOSO": SE PREOCUPA APENAS SOBRE A DIVISÃO ATUAL

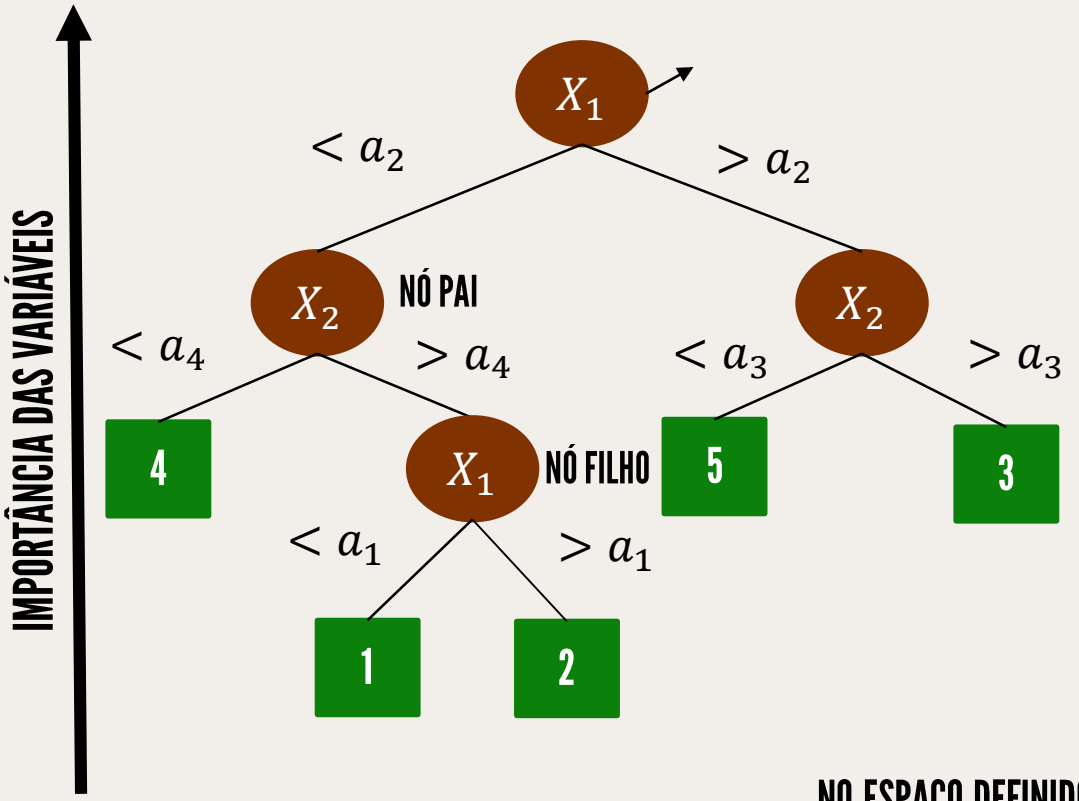
REPRESENTAÇÃO DE UMA ÁRVORE DE DECISÃO

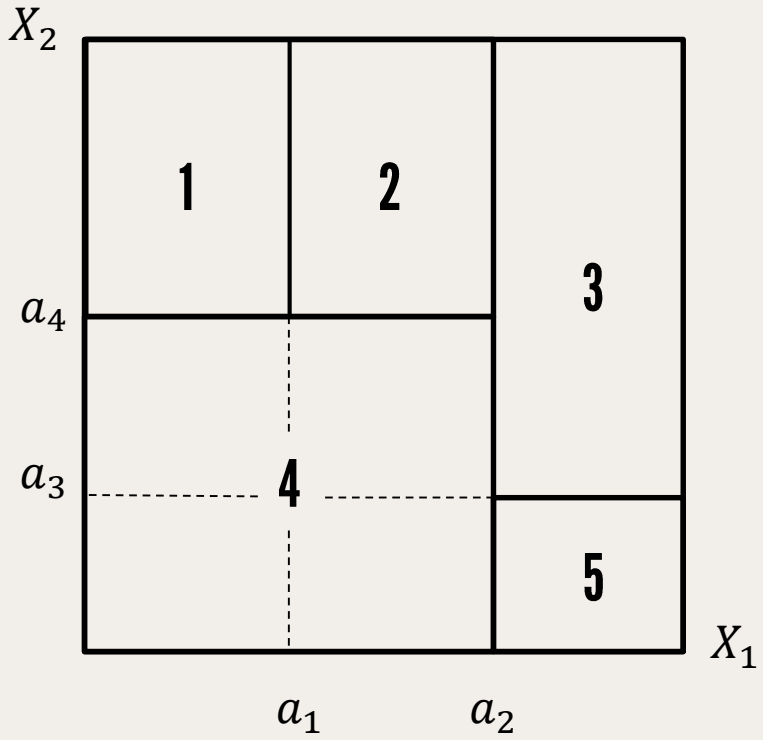




CADA PERCURSO DA ÁRVORE, DA RAIZ À FOLHA CORRESPONDE UMA REGRA

REPRESENTAÇÃO DE UMA ÁRVORE DE DECISÃO





NO ESPAÇO DEFINIDO PELOS ATRIBUTOS, CADA FOLHA CORRESPONDE A UM HIPER-RETÂNGULO, EM QUE A INTERSEÇÃO DESTES É VAZIA E A UNIÃO É TODO O ESPAÇO.

ÁRVORES DE REGRESSÃO × ÁRVORES DE CLASSIFICAÇÃO

ÁRVORES DE CLASSIFICAÇÃO	ÁRVORES DE REGRESSÃO
▪ VARIÁVEL DEPENDENTE CATEGÓRICA. ▪ O VALOR (CLASSE) OBTIDO PELO NÓ TERMINAL NOS DADOS DE TREINAMENTO É O MODA DAS OBSERVAÇÕES QUE CAEM NESSA REGIÃO.	▪ VARIÁVEL DEPENDENTE CONTÍNUA. ▪ O VALOR OBTIDO PELOS NÓS TERMINAIS NOS DADOS DE TREINAMENTO É A RESPOSTA MÉDIA DA OBSERVAÇÃO QUE CAI NESSA REGIÃO.

CONSTRUÇÃO DE UMA ÁRVORE DE DECISÃO

ESCOLHER UM ATRIBUTO

ESTENDER A ÁRVORE ADICIONANDO RAMOS DE ACORDO COM A PARTIÇÃO

PASSAR OS EXEMPLOS PARA OS NÓS (TENDO EM CONTA O VALOR DO ATRIBUTO ESCOLHIDO).

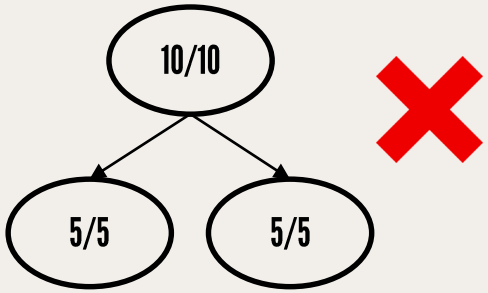
AVALIAR O CRITÉRIO DE PARADA

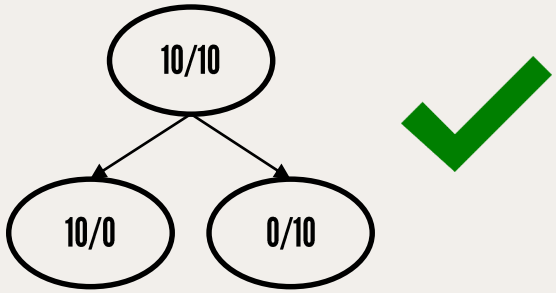
OBJETIVO: EXECUTAR UMA SEQUÊNCIA DE DIVISÕES DE CIMA PARA BAIXO A FIM DE CRIAR NÓS TERMINAIS (FOLHAS) EM QUE AS CLASSES ESTÃO BEM SEPARADAS (CLASSIFICAÇÃO) OU O ERRO QUADRÁTICO MÉDIO É PEQUENO (REGRESSÃO).

CRITÉRIOS PARA ESCOLHA DO ATRIBUTO

- O critério utilizado para realizar as partições é o da utilidade do atributo para a classificação/regressão.

- Na classificação, p. ex.:
  - Uma divisão que mantêm as proporções de classes em todas as partições é inútil.
  - Uma divisão onde em cada partição todos os exemplos são da mesma classe tem utilidade máxima.





Classificação

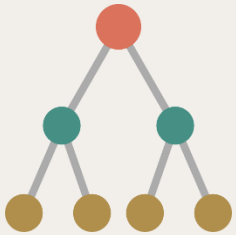
MEDIDAS DE PARTIÇÃO

- Maioria dos algoritmos de indução de árvores de decisão trabalha com funções de **divisão univariável**, ou seja, cada nó interno da árvore é dividido de acordo com um único atributo.

- Os critérios de seleção para a melhor divisão são baseados em diferentes medidas, tais como impureza, distância e dependência.

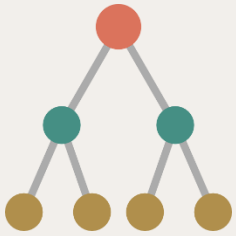
- Na maioria dos casos, nem todas as possíveis variáveis de entrada serão usadas para construir o modelo de árvore de decisão e, em alguns casos, uma variável de entrada específica pode ser usada várias vezes em diferentes níveis da árvore de decisão.

MEDIDAS DE PARTIÇÃO



CLASSIFICAÇÃO:

- GANHO DE INFORMAÇÃO
- ÍNDICE GINI



REGRESSÃO:

- SDR (DO INGLÊS *STANDARD DEVIATION REDUCTION*).

GANHO DE INFORMAÇÃO

Dado um conjunto de exemplos, qual atributo escolher para realizar a partição?

- O ganho de informação mede a redução da entropia causada pela partição dos exemplos de acordo com os valores do atributo.
- Entropia é uma medida da aleatoriedade (grau de pureza) de uma variável.
- A entropia de uma variável nominal *X*, com instâncias pertencentes à classe i, com probabilidade p_i, é dada por

$$Entropia\ (X) = -\sum_i p_i \log_2 p_i$$

A ENTROPIA É MÁXIMA (IGUAL A 1) QUANDO O CONJUNTO DE DADOS É HETEROGÊNEO.

GANHO DE INFORMAÇÃO

$$Ganho\ (Exs, Atr) = \underbrace{Entropia\ (Exs)}_{\text{Nó pai}} - \sum_{x \in P(Atr)} \underbrace{\frac{\#Exs_x}{\#Exs} Entropia(Exs_x)}_{\text{Nós filhos}}$$

Atr: atributo;
Entropia (Exs): entropia dos exemplos;
P(Atr): conjunto dos valores que Atr pode assumir;
x: elemento deste conjunto;
Exs_x: subconjunto de Exs formado pelos dados em que Atr = x;
Entropia (Exs_x): entropia que se obtém ao particionar Exs em função do atributo Atr.

A construção de uma árvore de classificação é guiada pelo objetivo de diminuir a entropia.

EXEMPLO – GI (ATRIBUTOS NOMINAIS)

DIA	APARÊNCIA	JOGAR TÊNIS
1	Ensolarado	Não
2	Ensolarado	Não
3	Nublado	Sim
4	Chuva	Sim
5	Chuva	Sim
6	Chuva	Não
7	Nublado	Sim
8	Ensolarado	Não
9	Ensolarado	Sim
10	Chuva	Sim
11	Ensolarado	Sim
12	Nublado	Sim
13	Nublado	Sim
14	Chuva	Não

	Sol	Nublado	Chuva
Sim	2	4	3
Não	3	0	2

$$Ganho\ (Exs, Atr) =$$

$$= Entropia\ (Exs) - \sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} Entropia(Exs_x)$$

EXEMPLO – GI (ATRIBUTOS NOMINAIS)

DIA	APARÊNCIA	JOGAR TÊNIS
1	Ensolarado	Não
2	Ensolarado	Não
3	Nublado	Sim
4	Chuva	Sim
5	Chuva	Sim
6	Chuva	Não
7	Nublado	Sim
8	Ensolarado	Não
9	Ensolarado	Sim
10	Chuva	Sim
11	Ensolarado	Sim
12	Nublado	Sim
13	Nublado	Sim
14	Chuva	Não

	Sol	Nublado	Chuva
Sim	2	4	3
Não	3	0	2

$$Ganho\ (Exs, Atr) =$$

$$= \underbrace{Entropia\ (Exs)} - \sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} Entropia(Exs_x)$$

$$Entropia\ (Jogar\ Tênis) = -\sum_i p_i \log_2 p_i \qquad \text{2 classes}$$

$$Entropia\ (Jogar\ Tênis) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14} = 0{,}940$$

EXEMPLO – GI (ATRIBUTOS NOMINAIS)

DIA	APARÊNCIA	JOGAR TÊNIS
1	Ensolarado	Não
2	Ensolarado	Não
3	Nublado	Sim
4	Chuva	Sim
5	Chuva	Sim
6	Chuva	Não
7	Nublado	Sim
8	Ensolarado	Não
9	Ensolarado	Sim
10	Chuva	Sim
11	Ensolarado	Sim
12	Nublado	Sim
13	Nublado	Sim
14	Chuva	Não

	Sol	Nublado	Chuva
Sim	2	4	3
Não	3	0	2

$$Ganho\ (Exs, Atr) =$$

$$= Entropia\ (Exs)\ - \underbrace{\sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} Entropia(Exs_x)}$$

$$\frac{\#Exs_s}{\#Exs} Entropia(Exs_s) + \frac{\#Exs_n}{\#Exs} Entropia(Exs_n) + \frac{\#Exs_c}{\#Exs} Entropia(Exs_c)$$

$$\frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{4}{14}\left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) + \frac{5}{14}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$$

$$= 0{,}693$$

EXEMPLO – GI (ATRIBUTOS NOMINAIS)

DIA	APARÊNCIA	JOGAR TÊNIS
1	Ensolarado	Não
2	Ensolarado	Não
3	Nublado	Sim
4	Chuva	Sim
5	Chuva	Sim
6	Chuva	Não
7	Nublado	Sim
8	Ensolarado	Não
9	Ensolarado	Sim
10	Chuva	Sim
11	Ensolarado	Sim
12	Nublado	Sim
13	Nublado	Sim
14	Chuva	Não

	Sol	Nublado	Chuva
Sim	2	4	3
Não	3	0	2

$$Ganho\ (Exs, Atr) =$$

$$= Entropia\ (Exs)\ - \sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} Entropia(Exs_x)$$

$$Ganho\ (Exs, Atr) = 0{,}940 - 0{,}693 = 0{,}247$$

GI (ATRIBUTOS NUMÉRICOS)

- No caso de atributos numéricos, o teste produz uma partição binária do conjunto de exemplos:
  - Exemplos em que o valor_do_atributo < ponto_referência
  - Exemplos em que o valor_do_atributo >= ponto_referência
- Escolha do ponto de referência:
  - Ordenar os exemplos por ordem crescente dos valores do atributo numérico.
  - Qualquer ponto intermediário entre dois valores diferentes (da classe) e consecutivos dos valores observados no conjunto de treinamento pode ser utilizado como possível ponto de referência.
  - É usual considerar o valor médio entre dois valores diferentes e consecutivos.

EXEMPLO – GI (ATRIBUTOS NUMÉRICOS)

TEMPERATURA	JOGAR TÊNIS	
64	Sim	←
65	Não	
68	Sim	←
69	Sim	
70	Sim	←
71	Não	
72	Sim	←
72	Sim	
75	Sim	←
75	Não	
80	Sim	←
81	Sim	
83	Sim	←
85	Não	

$Temperatura < 70{,}5$

$Temperatura \geq 70{,}5$

	$< 70{,}5$	$\geq 70{,}5$
Sim	4	5
Não	1	4

$Ganho\ (Exs, Atr)$

GANHO DE INFORMAÇÃO

DEFINIDO O ATRIBUTO COM MAIOR GANHO DE INFORMAÇÃO, ESTE SERÁ UTILIZADO PARA INICIAR A PARTIÇÃO.

REPETE-SE O PROCESSO DE AVALIAÇÃO DOS ATRIBUTOS ATÉ O FINAL DA CONSTRUÇÃO DA ÁRVORE.

ÍNDICE GINI

- O índice Gini mede o grau de heterogeneidade dos dados → utilizado para medir a impureza de um nó.
- Este índice num determinado nó é dado por:

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

p_i é a frequência relativa de cada classe em cada nó e c é o número de classes.

- Quando este índice é igual a zero, o nó é puro. Quando se aproxima de 1, o nó é impuro.

$$Índice\ Gini_{Atr} = Índice\ Gini(Exs) - \sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} Índice\ Gini(Exs_x)$$

P(Atr): conjunto dos valores que Atr pode assumir;
x: elemento deste conjunto;
Exs_x: subconjunto de Exs formado pelos dados em que Atr = x.

EXEMPLO – GINI (ATRIBUTOS NOMINAIS)

DIA	APARÊNCIA	JOGAR TÊNIS
1	Ensolarado	Não
2	Ensolarado	Não
3	Nublado	Sim
4	Chuva	Sim
5	Chuva	Sim
6	Chuva	Não
7	Nublado	Sim
8	Ensolarado	Não
9	Ensolarado	Sim
10	Chuva	Sim
11	Ensolarado	Sim
12	Nublado	Sim
13	Nublado	Sim
14	Chuva	Não

	Sol	Nublado	Chuva
Sim	2	4	3
Não	3	0	2

$$\text{Índice Gini}_{Atr} = \underbrace{\text{Índice Gini}(Exs)}_{\text{Nó pai}} - \underbrace{\sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} \text{Índice Gini}(Exs_x)}_{\text{Nós filhos}}$$

EXEMPLO – GINI (ATRIBUTOS NOMINAIS)

DIA	APARÊNCIA	JOGAR TÊNIS
1	Ensolarado	Não
2	Ensolarado	Não
3	Nublado	Sim
4	Chuva	Sim
5	Chuva	Sim
6	Chuva	Não
7	Nublado	Sim
8	Ensolarado	Não
9	Ensolarado	Sim
10	Chuva	Sim
11	Ensolarado	Sim
12	Nublado	Sim
13	Nublado	Sim
14	Chuva	Não

	Sol	Nublado	Chuva
Sim	2	4	3
Não	3	0	2

$$\text{Índice Gini}_{Atr} = \underbrace{\text{Índice Gini}(Exs)} - \sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} \text{Índice Gini}(Exs_x)$$

$$\text{Índice Gini}(Exs) = 1 - \left(\frac{5}{14}\right)^2 - \left(\frac{9}{14}\right)^2 = 0{,}4592$$

EXEMPLO – GINI (ATRIBUTOS NOMINAIS)

DIA	APARÊNCIA	JOGAR TÊNIS
1	Ensolarado	Não
2	Ensolarado	Não
3	Nublado	Sim
4	Chuva	Sim
5	Chuva	Sim
6	Chuva	Não
7	Nublado	Sim
8	Ensolarado	Não
9	Ensolarado	Sim
10	Chuva	Sim
11	Ensolarado	Sim
12	Nublado	Sim
13	Nublado	Sim
14	Chuva	Não

	Sol	Nublado	Chuva
Sim	2	4	3
Não	3	0	2

$$\text{Índice Gini}_{Atr} = \text{Índice Gini}(Exs) - \sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} \text{Índice Gini}(Exs_x)$$

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

$$\sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} \text{Índice Gini}(Exs_x) = \frac{5}{14}\left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) +$$

$$+\frac{4}{14}\left(1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2\right) + \frac{5}{14}\left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) = 0{,}3429$$

EXEMPLO – GINI (ATRIBUTOS NOMINAIS)

DIA	APARÊNCIA	JOGAR TÊNIS
1	Ensolarado	Não
2	Ensolarado	Não
3	Nublado	Sim
4	Chuva	Sim
5	Chuva	Sim
6	Chuva	Não
7	Nublado	Sim
8	Ensolarado	Não
9	Ensolarado	Sim
10	Chuva	Sim
11	Ensolarado	Sim
12	Nublado	Sim
13	Nublado	Sim
14	Chuva	Não

	Sol	Nublado	Chuva
Sim	2	4	3
Não	3	0	2

$$\text{Índice } Gini_{Atr} = \text{Índice } Gini(Exs) - \sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} \text{Índice } Gini(Exs_x)$$

$$\text{Índice } Gini_{Atr} = 0{,}4592 - 0{,}3429 = 0{,}1163$$

ÍNDICE GINI (ATRIBUTOS CONTÍNUOS)

O MESMO PROCEDIMENTO PARA O GANHO DE INFORMAÇÃO É REALIZADO PARA ATRIBUTOS NUMÉRICOS.

DEFINIDO O ATRIBUTO COM MAIOR ÍNDICE GINI, REPETE-SE O PROCESSO PARA CADA PARTIÇÃO.

SDR (*STANDARD DEVIATION REDUCTION*)

- Assuma um conjunto de exemplos. O desvio padrão da variável alvo, y, é dada pela expressão

$$sd(Exs) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- O desvio padrão de y em cada partição é sempre menor ou igual ao desvio padrão de y antes da divisão. Podemos estimar essa redução como

$$SDR_{Atr} = sd(Exs) - \sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} sd\,(Exs_x)$$

P(Atr): conjunto dos valores que Atr pode assumir;
x: elemento deste conjunto;
Exs_x: subconjunto de Exs formado pelos dados em que Atr = x.

O teste que provoca uma maior redução no desvio padrão é escolhido como teste para o nó.

EXEMPLO – SDR

DIA	APARÊNCIA	HORAS JOGADAS
1	Ensolarado	45
2	Ensolarado	52
3	Nublado	46
4	Chuva	25
5	Chuva	30
6	Chuva	35
7	Nublado	43
8	Ensolarado	23
9	Ensolarado	46
10	Chuva	38
11	Ensolarado	30
12	Nublado	52
13	Nublado	44
14	Chuva	48

$$SDR_{Atr} = sd(Exs) - \sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} sd\,(Exs_x)$$

$$sd(Exs) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 9{,}32 \qquad \bar{y} = 39{,}8$$

$$\sum_{x \in P(Atr)} \frac{\#Exs_x}{\#Exs} sd\,(Exs_x) = \frac{\#Exs_c}{\#Exs} sd\,(Exs_c) + \frac{\#Exs_n}{\#Exs} sd\,(Exs_n) + \frac{\#Exs_s}{\#Exs} sd\,(Exs_s)$$

$$= \frac{5}{14} sd\,(Exs_c) + \frac{4}{14} sd\,(Exs_n) + \frac{5}{14} sd\,(Exs_s) = 7{,}66$$

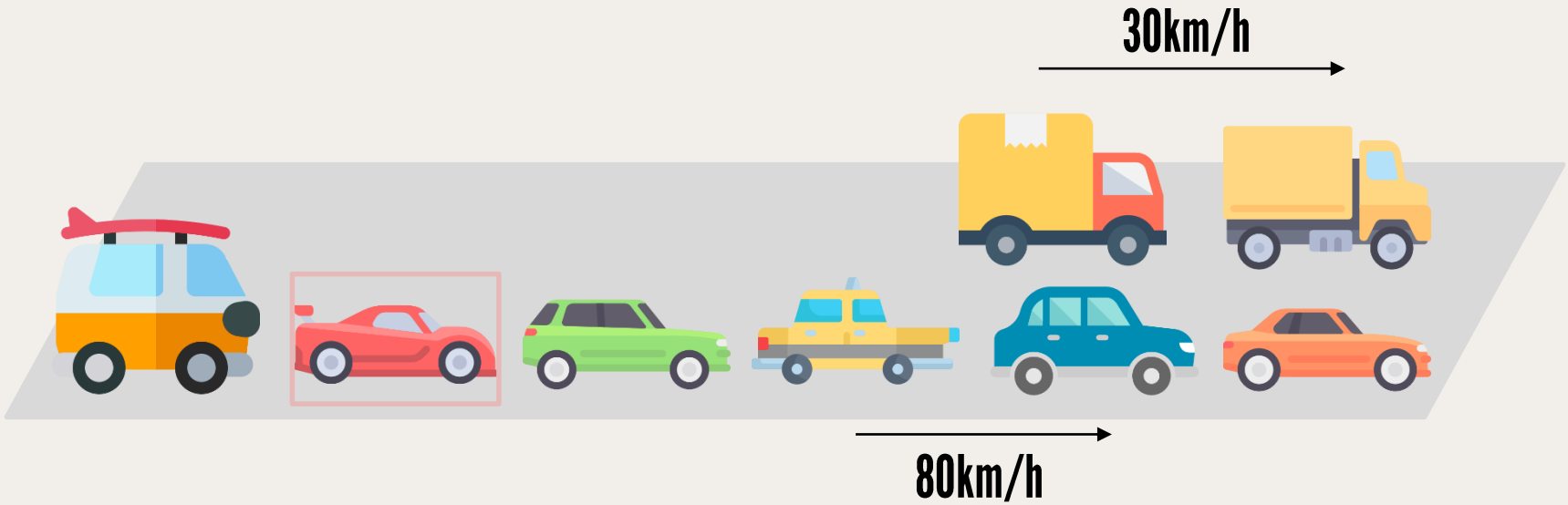
$$SDR_{Atr} = 9{,}32 - 7{,}66 = 1{,}66$$

- NÚMERO MÍNIMO DE OBSERVAÇÕES EM UM NÓ A SER CONSIDERADO PARA DIVISÃO
- NÚMERO MÍNIMO DE OBSERVAÇÕES EM UM NÓ TERMINAL (FOLHA)
- PROFUNDIDADE MÁXIMA DA ÁRVORE (PROFUNDIDADE VERTICAL)
- NÚMERO MÁXIMO DE NÓS TERMINAIS (PODE SER DEFINIDO NO LUGAR DA PROFUNDIDADE)
- MEDIDA DE PARTIÇÃO MENOR QUE UM VALOR PRÉ-DEFINIDO

CRITÉRIOS DE PARADA

PODA

Relembrando: algoritmo guloso → verifica a melhor divisão instantaneamente e avançará até que uma das condições de parada especificadas seja atingida.



Uma árvore de decisão com critérios de parada não verá o caminhão à frente e adotará uma abordagem gananciosa ao virar à esquerda. Por outro lado, se usarmos a poda, olhamos alguns passos à frente e fazemos uma escolha.

PODA

- Métodos de poda da árvore são utilizados para detectar e excluir ramos e sub-árvores com o objetivo de melhorar a taxa de acerto do modelo para novos exemplos.
- A árvore podada se torna mais simples, facilitando a sua interpretabilidade por parte do usuário.
- Pré-poda (critério de parada);
  - A pré-poda é mais rápida, porém menos eficiente que a pós-poda pelo fato do risco de interromper o crescimento da árvore ao selecionar uma árvore sub-ótima.
- Pós-poda.

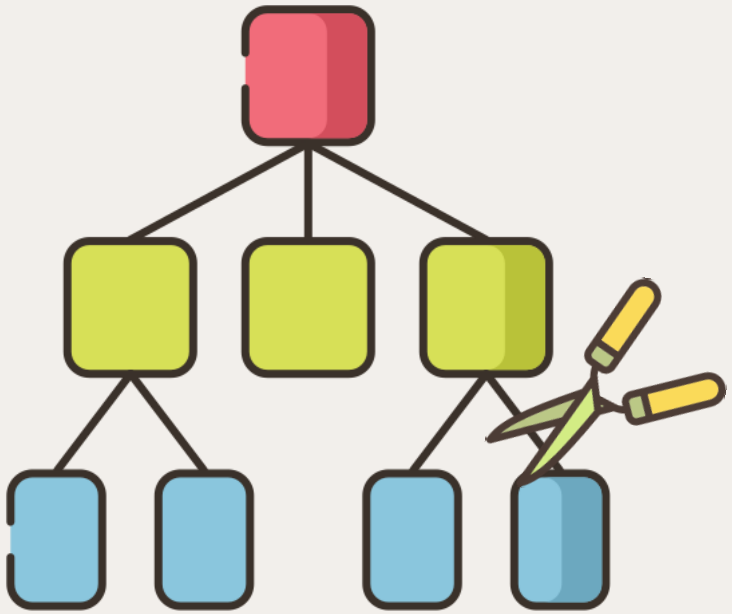
PÓS-PODA

- A pós-poda busca encontrar o tamanho adequado de uma árvore após a árvore ser induzida completamente.
  - É avaliada a confiabilidade de cada uma de suas sub-árvores, podando os ramos considerados não confiáveis.
- Dentre os métodos de pós-poda existentes, destacam-se:
  - Redução de erros (*Reduced Error Pruning*);
  - Custo-complexidade (*Cost-Complexity Pruning*);
  - Erro pessimista (*Pessimistic Error Pruning*);
  - Valor crítico (*Critical Value Pruning*);
  - Erro mínimo (*Minimum Error Pruning*);
  - Poda por estimativa de erro (*Error-Based Pruning*).

REDUÇÃO DE ERROS (REDUCED ERROR PRUNING)

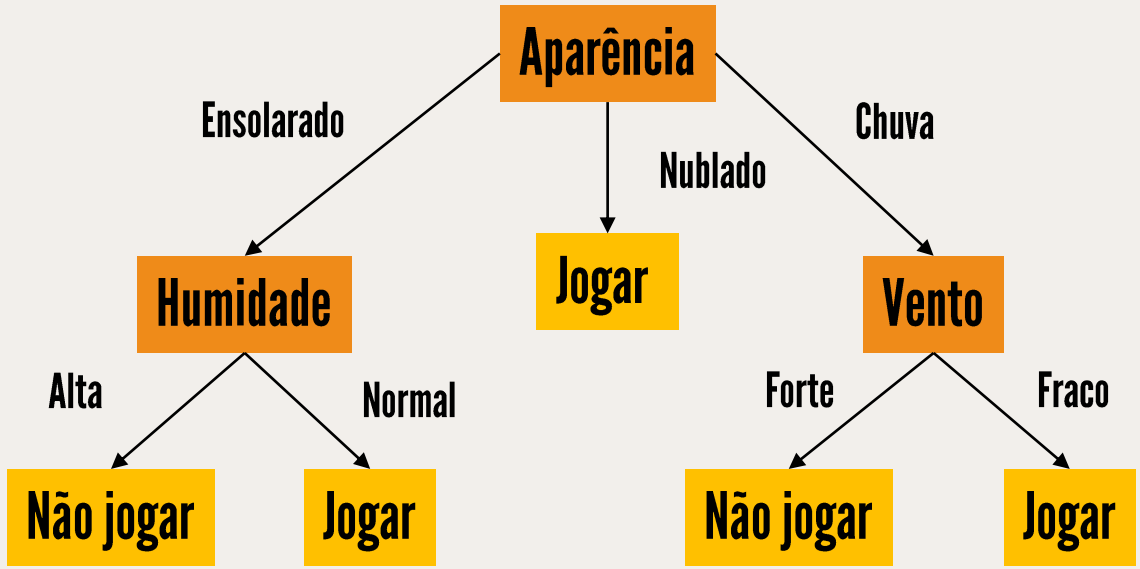
MÉTODO SIMPLES E RÁPIDO

SEGUE A ORIENTAÇÃO *BOTTOM-UP* E NECESSITA DE UM CONJUNTO DE VALIDAÇÃO PARA O PROCESSO DE PODA



REDUÇÃO DE ERROS (REDUCED ERROR PRUNING)

Particione os dados de treinamento em conjuntos “crescimento” e “validação”.
Construa uma árvore completa a partir dos dados de “crescimento”.

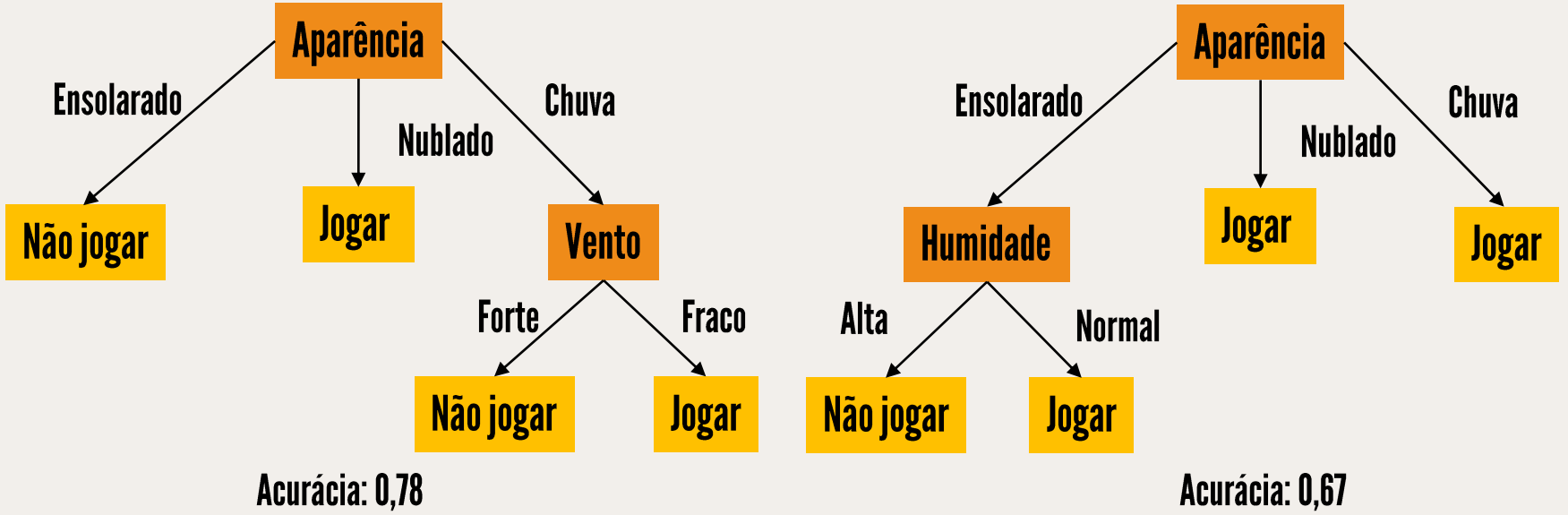


REDUÇÃO DE ERROS (REDUCED ERROR PRUNING)

Para cada nó não-folha, n, na árvore, faça:

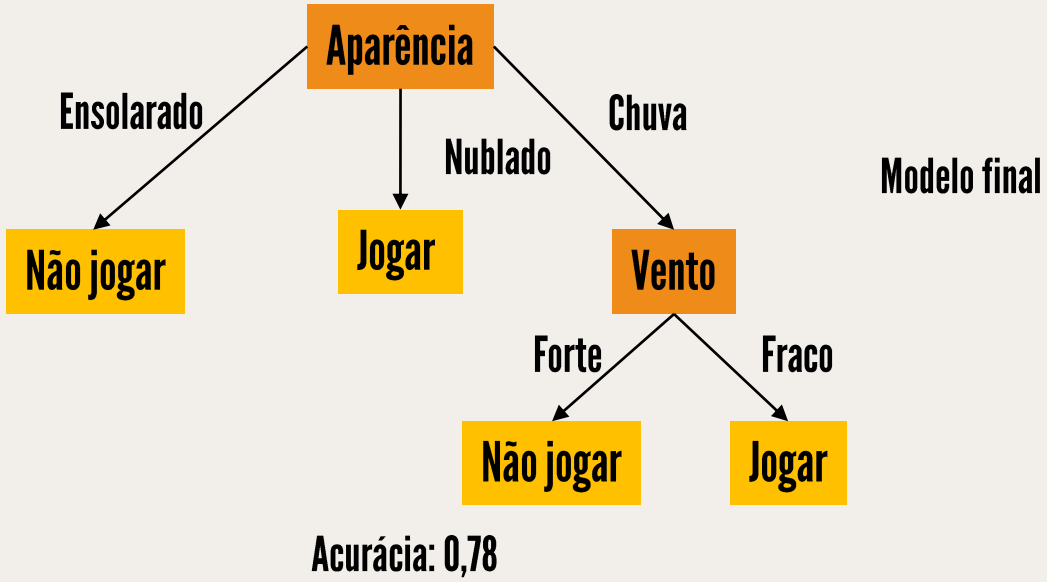
Remova temporariamente a sub-árvore abaixo de n e substitua-a por uma folha rotulada com a classe majoritária atual nesse nó.

Meça e registre a acurácia da árvore podada no conjunto de validação.



REDUÇÃO DE ERROS (REDUCED ERROR PRUNING)

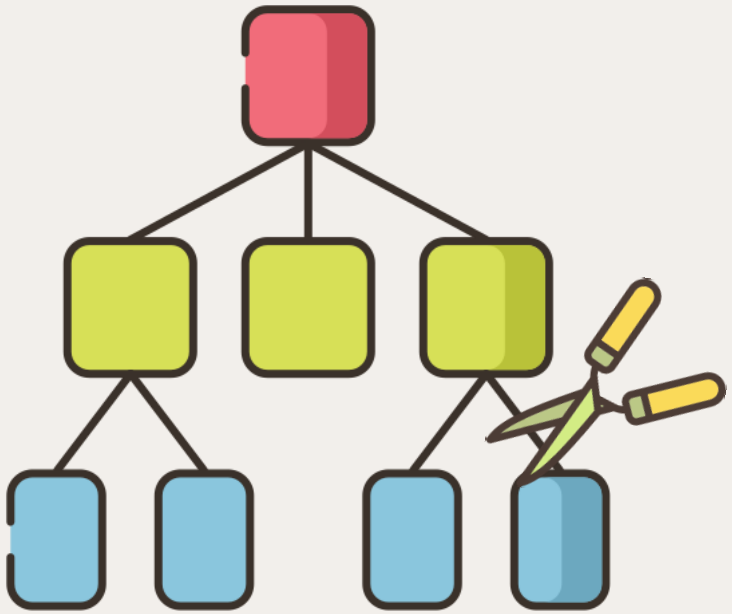
Remova permanentemente o nó que resulta no maior aumento da acurácia no conjunto de validação.
Continue este processo até que a poda reduza a acurácia no conjunto de validação.



CUSTO-COMPLEXIDADE (COST-COMPLEXITY PRUNING)

MÉTODO COMPOSTO POR DUAS ETAPAS

MAIS COMPLEXO QUE O DE REDUÇÃO DE ERROS



CUSTO-COMPLEXIDADE (COST-COMPLEXITY PRUNING)

1. Uma sequência de árvores T_0, T_1, ..., T_K é construída com os dados de treinamento.
   T_0 é a árvore original antes da poda e T_K é a árvore raiz (só com o nó raiz).

2. É selecionada a melhor árvore dessa sequência, levando-se em consideração o custo estimado dos erros/desvios de classificação/regressão e a complexidade (medida em número de folhas) de cada uma dessas árvores.

CUSTO-COMPLEXIDADE (COST-COMPLEXITY PRUNING)

- O primeiro passo é construir uma árvore suficientemente grande T_{max}.
  - T_{max} não precisa ser expandida de forma exaustiva, basta ser suficientemente grande. Para isso, basta estabelecer um número mínimo de exemplos por folha e adotá-lo como critério de parada.
- Para qualquer sub-árvore $T \preccurlyeq T_{max}$, defina sua complexidade como $|\bar{T}|$, o número de folhas em T. Seja $\alpha > 0$ um número real denominado parâmetro de complexidade e defina a medida de custo-complexidade $R_\alpha(T)$ como

$$R_\alpha(T) = R(T) + \alpha|\bar{T}|$$

CUSTO DE ERRO COMPLEXIDADE

- O problema central do método é encontrar, para cada valor de α, a sub-árvore $T(\alpha) \preccurlyeq T_{max}$ que minimiza $R_\alpha(T)$.

ALGORITMOS

Dependendo do problema, um algoritmo pode ser mais eficiente que outro.

Dentre os algoritmos, tem-se:

- ID3;
- C4.5;
- C5;
- CART – Classification and Regression Trees;

...

ID3

- Algoritmo pioneiro em indução de árvores de decisão → recursivo e baseado em busca gulosa.
- Limitações:
  - Só lida com atributos categóricos não-ordinais.
    - Os atributos contínuos devem ser previamente discretizados.
  - Não apresenta nenhuma forma para tratar valores desconhecidos.
    - Necessário gastar um bom tempo com pré-processamento dos dados.
  - Não apresenta nenhum método de pós-poda.
- Utiliza o ganho de informação para selecionar a melhor divisão.
- Apenas problemas de classificação.

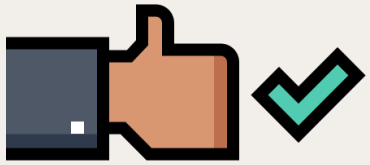
C4.5

- Evolução do ID3.
- Lida com atributos categóricos (ordinais ou não-ordinais) e atributos contínuos.
  - Atributos contínuos: define um limiar e então divide os exemplos de forma binária.
- Trata valores desconhecidos.
  - Permite que os valores desconhecidos para um determinado atributo sejam representados como '?', e o algoritmo trata esses valores de forma especial. Esses valores não são utilizados nos cálculos de ganho e entropia.
- Utiliza a medida de razão de ganho (ganho ponderado) para selecionar o atributo.
- Apresenta um método de pós-poda das árvores geradas.
- Algoritmo guloso, com estratégia "dividir para conquistar".
- Apenas problemas de classificação.

CART (CLASSIFICATION AND REGRESSION TREES)

- Classificação e regressão.
- As árvores são sempre binárias, questões simples do tipo "sim" ou "não".
  - Os nós que correspondem a atributos contínuos são representados por agrupamento de valores em dois conjuntos.
  - Dispõe de um tratamento especial para atributos ordenados e também permite a utilização de combinações lineares entre atributos (agrupamento de valores em vários conjuntos).
- Expande a árvore exaustivamente, realizando pós-poda por meio da redução do fator custo-complexidade.
- Utiliza o índice Gini.

VANTAGENS



FÁCIL DE ENTENDER

REPRESENTAÇÃO GRÁFICA INTUITIVA

ÚTIL EM EXPLORAÇÃO DE DADOS

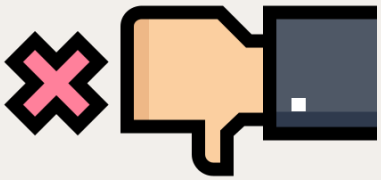
SELEÇÃO DE VARIÁVEIS

SENSIBILIDADE REDUZIDA A OUTLIERS

PODE MANIPULAR VARIÁVEIS NUMÉRICAS E CATEGÓRICAS

MÉTODO NÃO PARAMÉTRICO

DESVANTAGENS



OVERFITTING: RESTRIÇÕES SOBRE OS PARÂMETROS DO MODELO E PODA

INSTABILIDADE: PEQUENA MUDANÇA NOS DADOS PODE CAUSAR UMA GRANDE MUDANÇA NA ESTIMATIVA FINAL DA ÁRVORE

MENOR ACURÁCIA

RESUMO

- ALGORITMO DE APRENDIZADO SUPERVISIONADO USADO PRINCIPALMENTE EM PROBLEMAS DE CLASSIFICAÇÃO.
- FUNCIONA PARA VARIÁVEIS DE ENTRADA E SAÍDA CATEGÓRICAS E CONTÍNUAS.
- DIVISÃO DA AMOSTRA EM DOIS OU MAIS CONJUNTOS HOMOGÊNEOS COM BASE NO DIVISOR MAIS SIGNIFICATIVO DAS VARIÁVEIS DE ENTRADA.
- TIPO DE ÁRVORE MAIS COMUM: CART.