



ÁRVORES DE DECISÃO



Allan Dieguez

AI Researcher | Data Scientist

Director, Data Science

BAIN & COMPANY 

LinkedIn: *@allandieguez*

E-Mail: *allandieguez@gmail.com*



AGENDA

- **Mapeando Regras em Decisões**
Fundamentos de árvores de decisão
- **Árvores de Classificação**
Cálculo de pureza e entropia
- **Árvores de Regressão**
Cálculo de médias e resíduos
- **Regularização em Árvores**
Técnicas para evitar o *overfit*



MAPEANDO REGRAS EM DECISÕES

Fundamentos de árvores de decisão

x é um
número primo?

k = 0
i = 1

True

i <= **x**

False

x % **i** == 0

False

True

k += 1

i += 1

True

x
é primo

False

k == 2

x NÃO
é primo

ALGORITMO

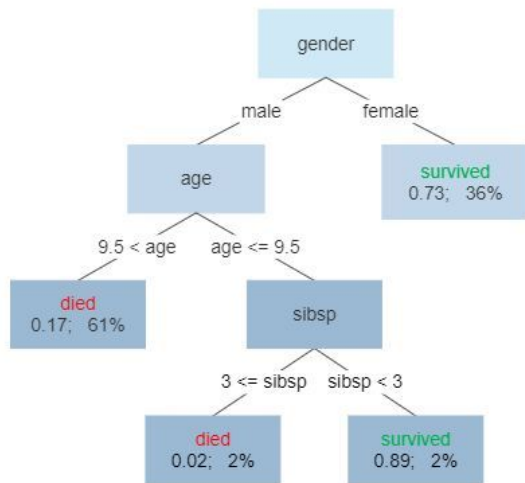
Sequência finita de regras que, aplicada a um número finito de dados, permite **solucionar classes semelhantes** de problemas.

Pode ser representado por um gráfico, chamado **fluxograma**.

As **decisões** do tipo **IF/ELSE** são representadas por **losangos**.

ÁRVORES DE DECISÃO

Survival of passengers on the Titanic



Fonte: [Wikipedia Commons](#)

DECISÕES BINÁRIAS

Para facilitar a **implementação**, árvores de decisão do tipo **CART*** são **sempre binárias**. Além de tornar mais simples a construção, **qualquer outro** tipo de partição pode ser representado por **um conjunto de partições binárias**.

INTERPRETAÇÃO

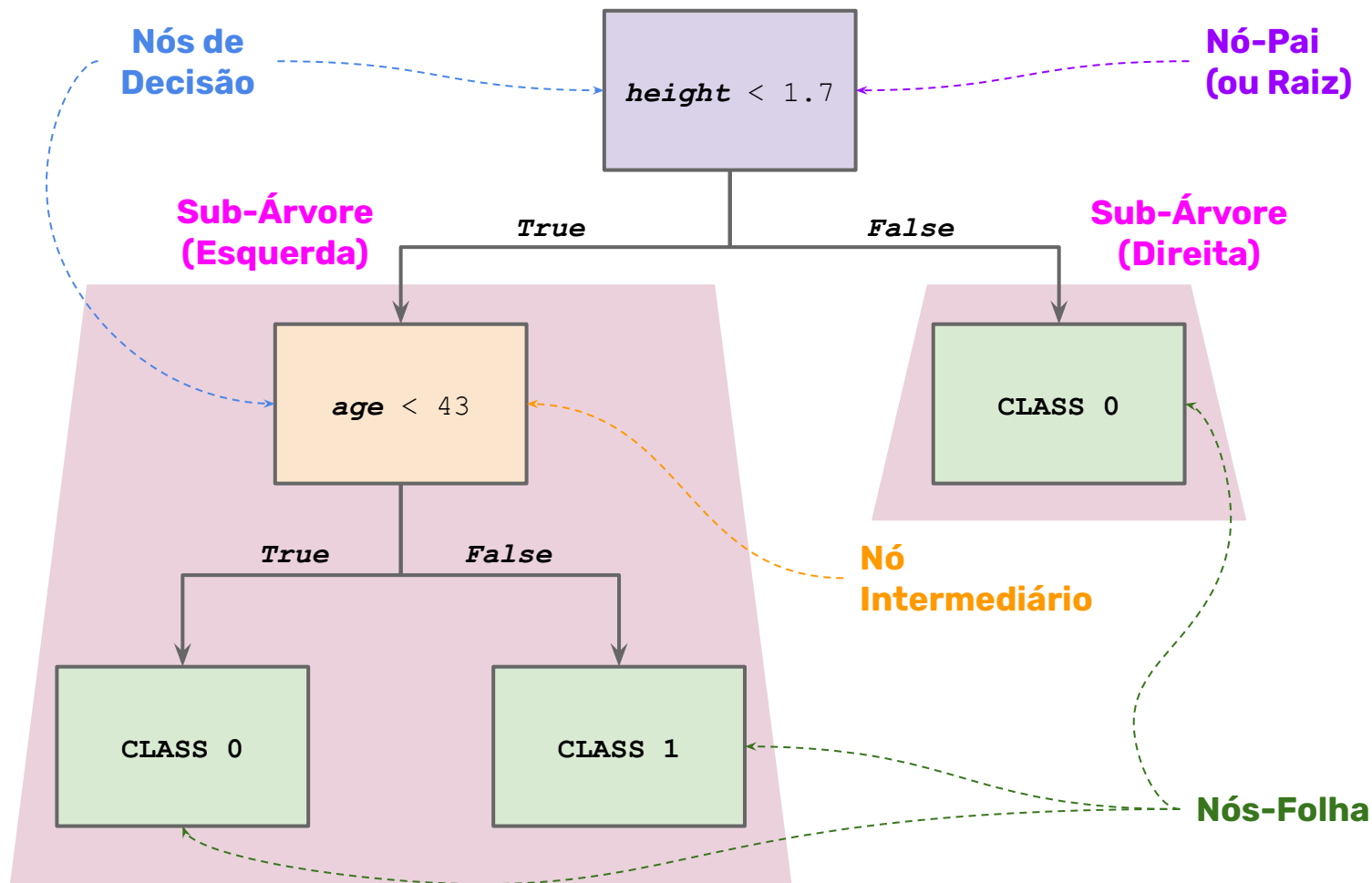
Cada **nó de decisão** é facilmente relacionado com uma decisão do tipo **IF/ELSE**. Os **nós-folha** representam uma **condição de parada**.

OUTROS TIPOS

Existem outros tipos de modelos **baseados em árvores não binárias**, como o **CHAID****, que permite que um nó-pai tenha mais de dois nós filhos

* **CART**: Classification And Regression Trees

** **CHAID**: Chi-square Automated Interaction Detection



VANTAGENS

DE USAR ÁRVORES DE DECISÃO

- **Fáceis de interpretar** como regras IF/ELSE
- Seleção de *features* é **parte da construção** da árvore, não sendo necessário aplicar outra metodologia
- **Escala bem** mesmo em datasets massivos

DESVANTAGENS

DE USAR ÁRVORES DE DECISÃO

- A **performance normalmente é mais baixa** quando comparadas com outros modelos mais complexos
- São modelos **muito sensíveis** a *outliers*, sofrendo bastante com **alta variância**



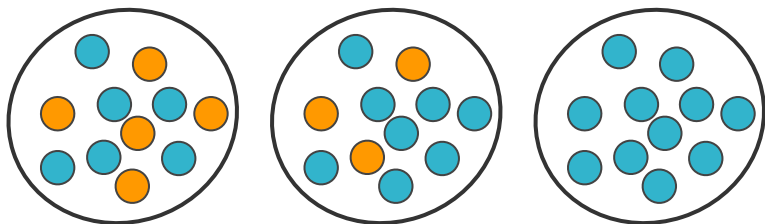
ÁRVORES DE CLASSIFICAÇÃO

Cálculo de pureza e entropia

ÁRVORES DE CLASSIFICAÇÃO

RUÍDO

INFORMAÇÃO



GRAU DE PUREZA

HOMOGENEIDADE DE UM NÓ

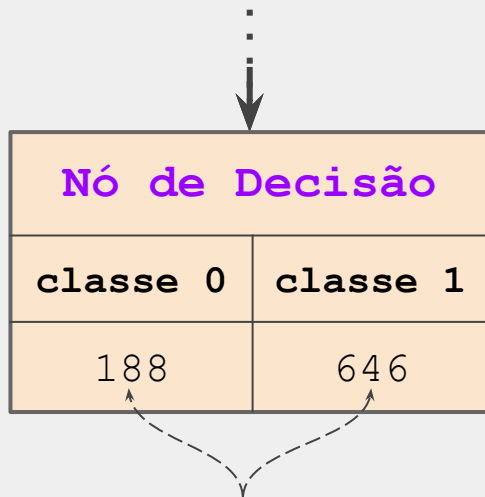
A **principal métrica de qualidade** para um nó da árvore de decisão é a **função de homogeneidade**. Na classificação, a homogeneidade é análoga à **pureza** de um conjunto de dados dentro de um nó.

RESPOSTA DA ÁRVORE

A resposta de um nó-folha é dada pela **classe de maior representação** dentro do nó. Quanto **mais homogênea** a composição do nó-folha, **maior a certeza** da resposta do classificador.

INCERTEZA DA RESPOSTA

Como a resposta do nó-folha é dada com base em um **conjunto de elementos**, pode-se calcular a **probabilidade** do nó pertencer a cada classe .



Quantidade de elementos **por classe**

$$\begin{aligned}
 N_{\text{Classe } 0} &= 188 \\
 N_{\text{Classe } 1} &= 646 \\
 N_{\text{Total}} &= 834
 \end{aligned}$$

$$P_{\text{Classe } 0} = N_{\text{Classe } 0} / N_{\text{Total}} = 0.23$$

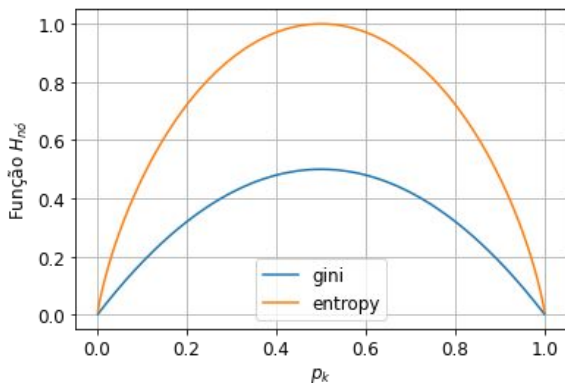
$$P_{\text{Classe } 1} = N_{\text{Classe } 1} / N_{\text{Total}} = 0.77$$

ELEMENTOS DO NÓ DE DECISÃO

Todo nó de uma **árvore de classificação** computa as **quantidades de elementos** de cada classe.

A **proporção** de elementos por classe é a base do cálculo dos **índices de impureza**.

MEDIDAS DE IMPUREZA



Fórmulas

Gini: $H_{\text{Gini}} = 1 - \sum_k p_k^2$

Entropia: $H_{\text{Entropy}} = -\sum_k p_k \log_2(p_k)$

MÉTRICAS MAIS UTILIZADAS

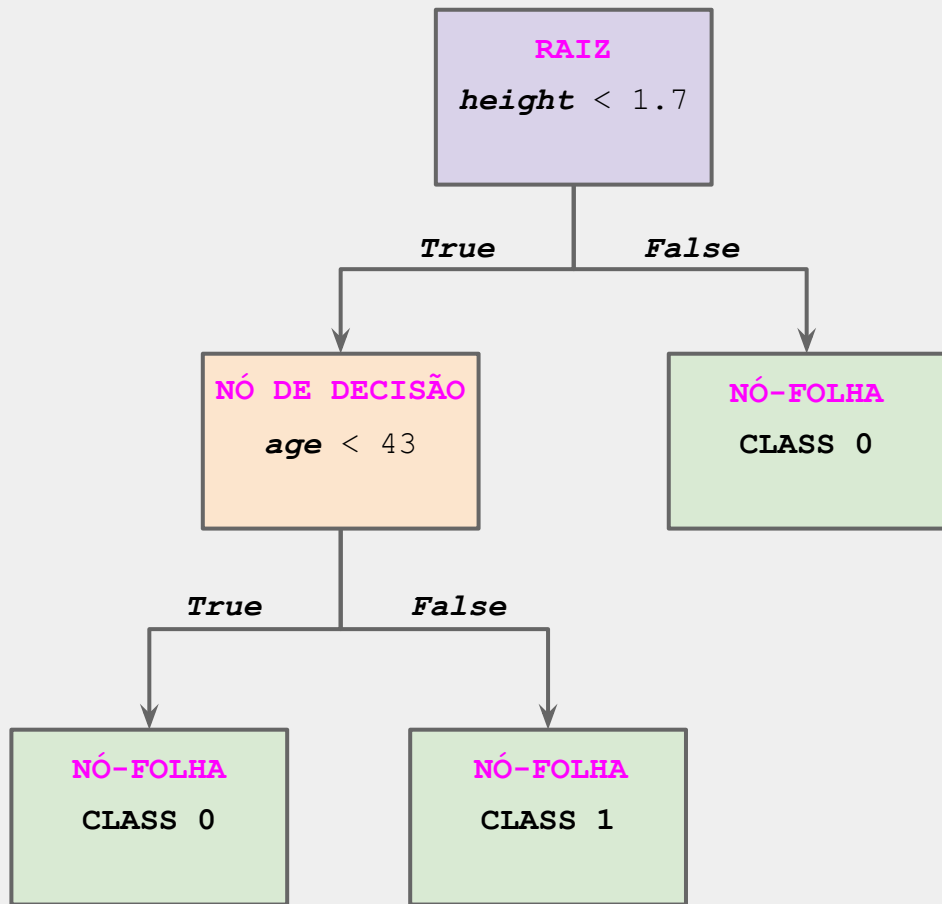
As métricas mais utilizadas para classificação são a de **Impureza de Gini** e a **Entropia**. Ambas seguem o princípio de que quanto **menor o valor, maior a qualidade** do nó.

GINI

A medida de **Impureza de Gini** é uma métrica que descreve a **probabilidade** de um elemento ser **aleatoriamente classificado de forma errada**. Um valor **zero** indica que o nó é puro, i.e. composto apenas por elementos de uma classe.

ENTROPIA

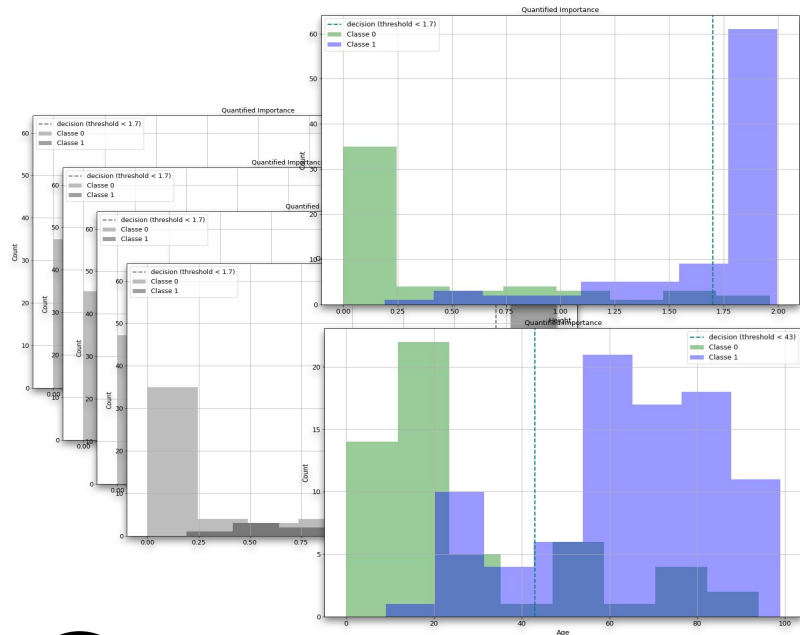
A **entropia** é uma métrica de **desordem** de um dado sistema. Um valor de **zero** indica que o nó é **completamente ordenado**, ou seja, não há dúvidas sobre a informação que contém.



CONSTRUÇÃO DA ÁRVORE

O treinamento do modelo consiste em **selecionar as features mais promissoras** em cada nó de decisão e avaliar a possibilidade de 1) criar um **nó-folha** ou 2) **continuar expandindo** a árvore.

O **primeiro passo**, porém, é a **escolha da Raiz** da árvore.



feature
candidata

Homogeneidade
(Função H_{Raiz})

height

0.606

age

0.877

gender

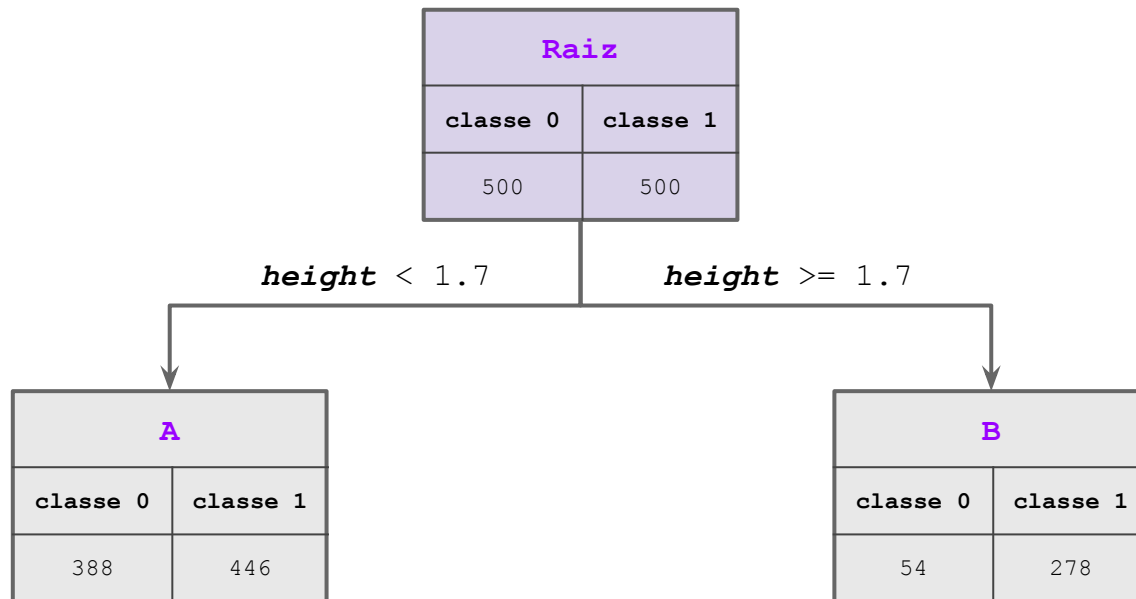
0.658

...

...

PASSO 1

Seleção da
feature mais
informativa

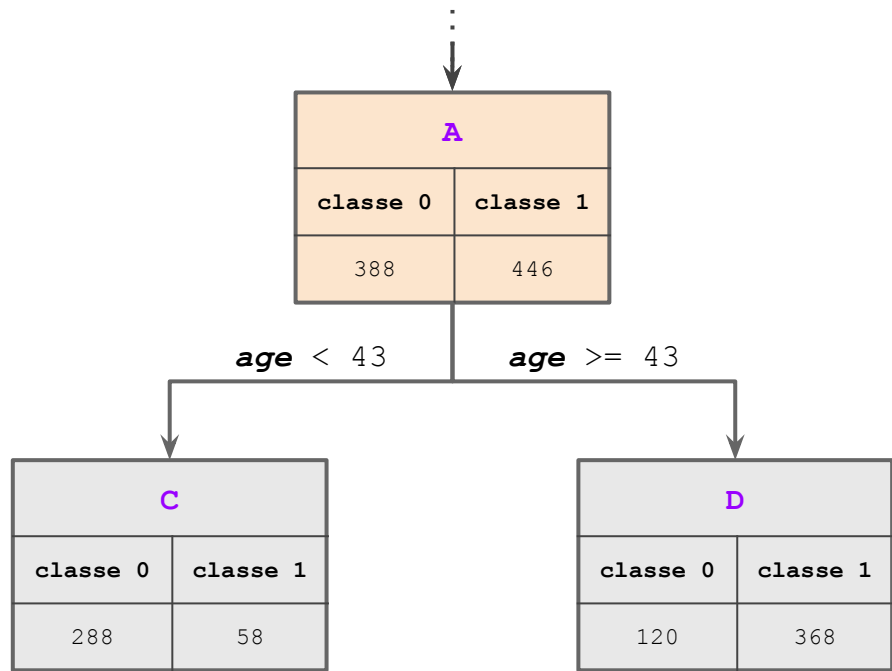


PASSO 1
Seleção da
feature mais
informativa

PASSO 2
Criação da
árvore com a
regra criada

T

<i>feature</i> candidata	Homogeneidade (Função H_A)
<i>height</i>	0.492
<i>age</i>	0.333
<i>gender</i>	0.613
...	...



PASSO 1
Seleção da
feature mais
informativa

PASSO 2
Criação da
árvore com a
regra criada

PASSO 3
Seleção de
feature no
nó-filho A

Racional:

Se $H_A > G_A$:
split é válido

Senão:
 Nó é Folha

Função G:

$$G_A = (N_{FE} \cdot H_{FE} + N_{FD} \cdot H_{FD}) / N_A$$

A	
classe 0	classe 1
388	446

$$N_A = 854$$

$$H_A = 0.498$$

C	
classe 0	classe 1
288	58

$$N_C = 346$$

$$H_C = 0.279$$

D	
classe 0	classe 1
120	368

$$N_D = 488$$

$$H_D = 0.371$$

$$G_A = (N_C \cdot H_C + N_D \cdot H_D) / N_A = 0.333$$

Então: $H_A > G_A$ É válido o *split*

PASSO 1
 Seleção da
feature mais
 informativa

PASSO 2
 Criação da
 árvore com a
 regra criada

PASSO 3
 Seleção de
feature no
 nó-filho A

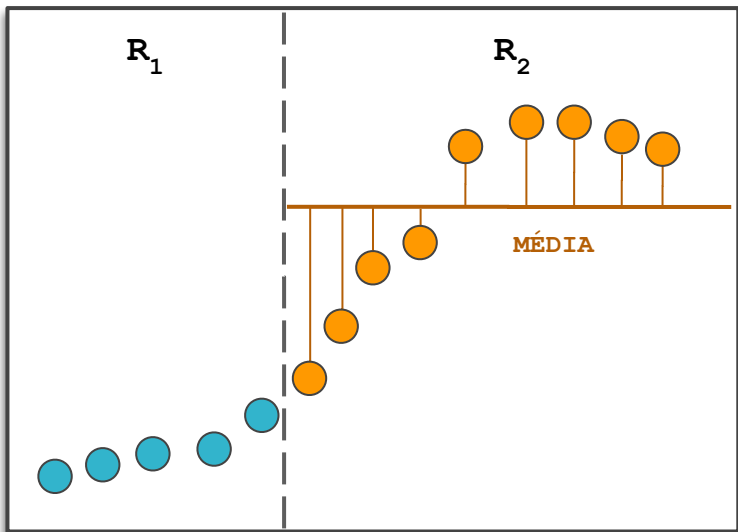
PASSO 4
 Decisão:
split ou
 nó-folha



ÁRVORES DE REGRESSÃO

Cálculo de médias e resíduos

ÁRVORES DE REGRESSÃO



HOMOGENEIDADE DE UM NÓ

A **função de homogeneidade** para a regressão está relacionada à **distância média** entre os valores dos elementos dentro do nó. Quanto **mais próximos** em valor são os elementos, **menor a distância** destes à média.

RESPOSTA DA ÁRVORE

A resposta de um nó-folha é dada pela **média dos valores** dos elementos que o compõem. Quanto **menor a distância** dos elementos à média, **mais assertiva** é a resposta do regressor.

INCERTEZA DA RESPOSTA

Como a resposta do nó-folha é dada com base em um **conjunto de elementos**, pode-se calcular a **variância** da resposta em torno da média.

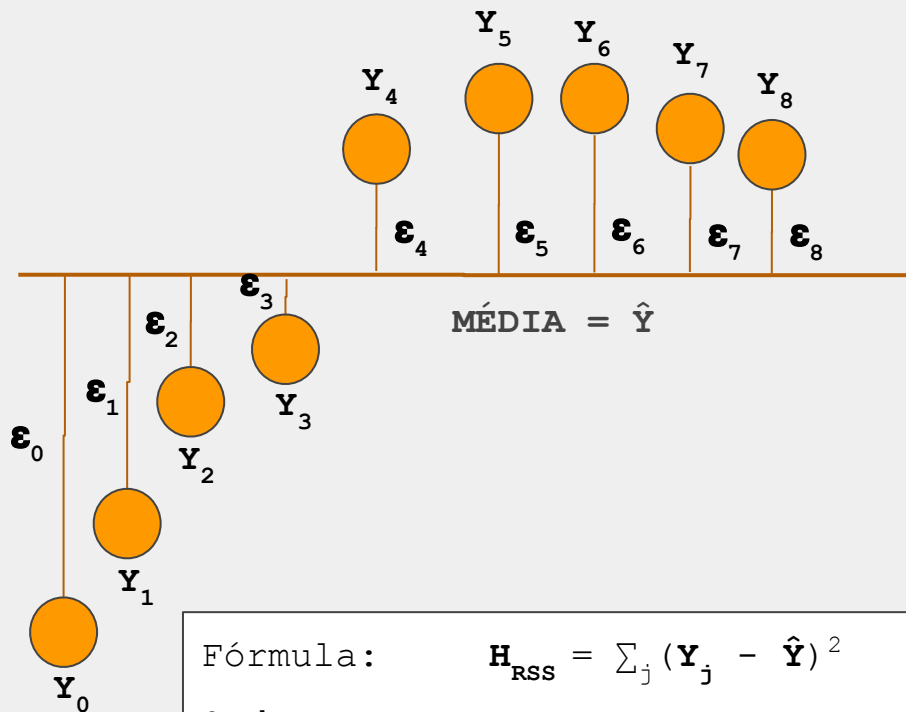


Nó de Decisão	
Número de Elementos	755.0
Média dos Valores	47.8
Métrica de Distância	1223.7

ELEMENTOS DO NÓ DE DECISÃO

Os componentes do nó de uma **árvore de regressão** descrevem, além do **número total** de elementos, a sua **média** e uma **métrica de distância** a essa média.

Essa métrica de distância é calculada sobre os **resíduos** da regressão. Por esse motivo, algumas vezes é referenciado como **erro** de predição.



Fórmula:
$$H_{RSS} = \sum_j (Y_j - \hat{Y})^2$$

Onde:

Y_j : valor do elemento j

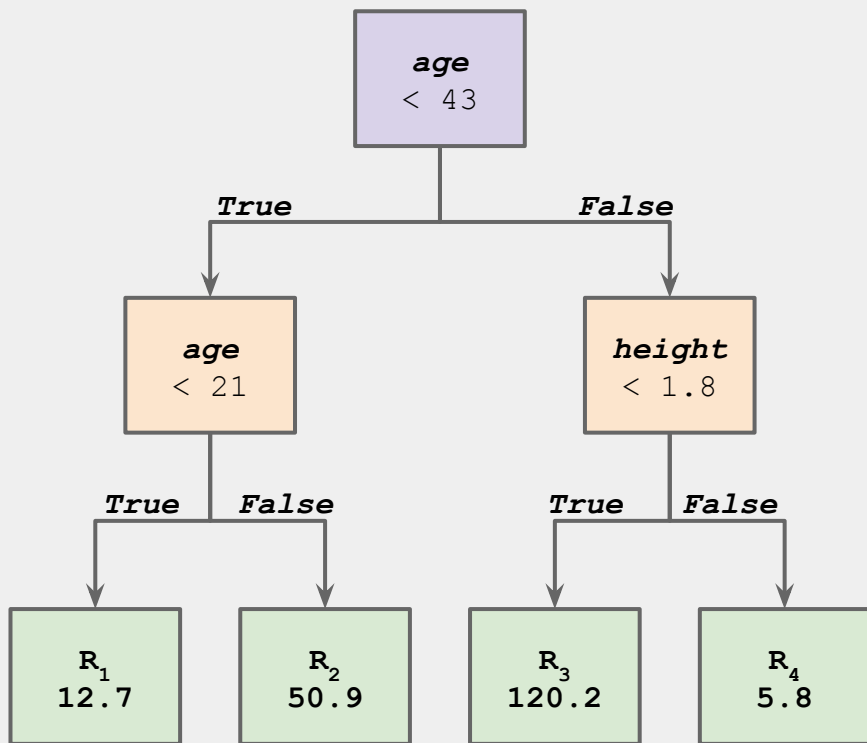
\hat{Y} : média dos elementos no nó

$\epsilon_j = Y_j - \hat{Y}$: resíduo do elemento j

MÉTRICA RSS

A **métrica mais utilizada** em árvores de regressão é o **RSS** (*Residual Sum of Squares*), um valor que sumariza de forma absoluta a **distância à média** de todos os elementos dentro de um nó.

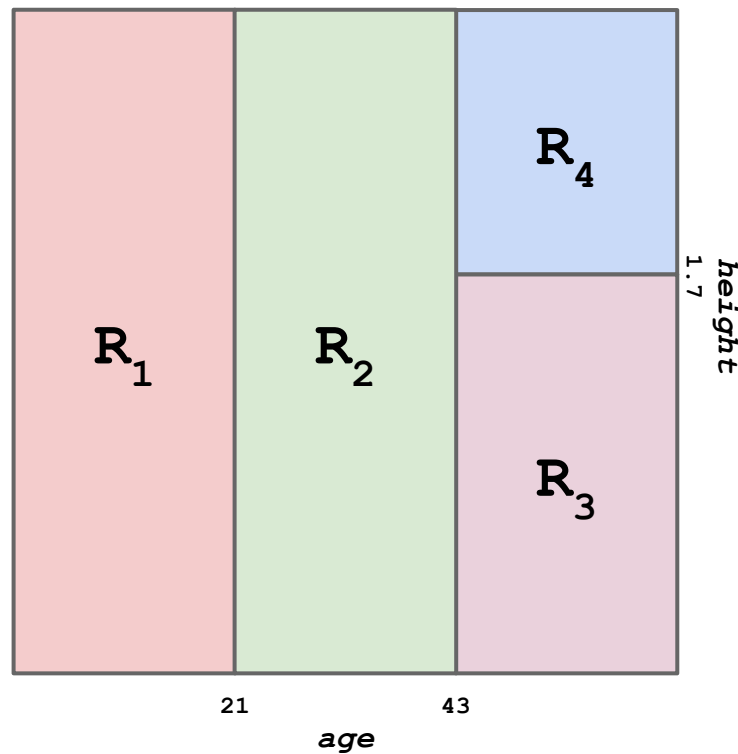
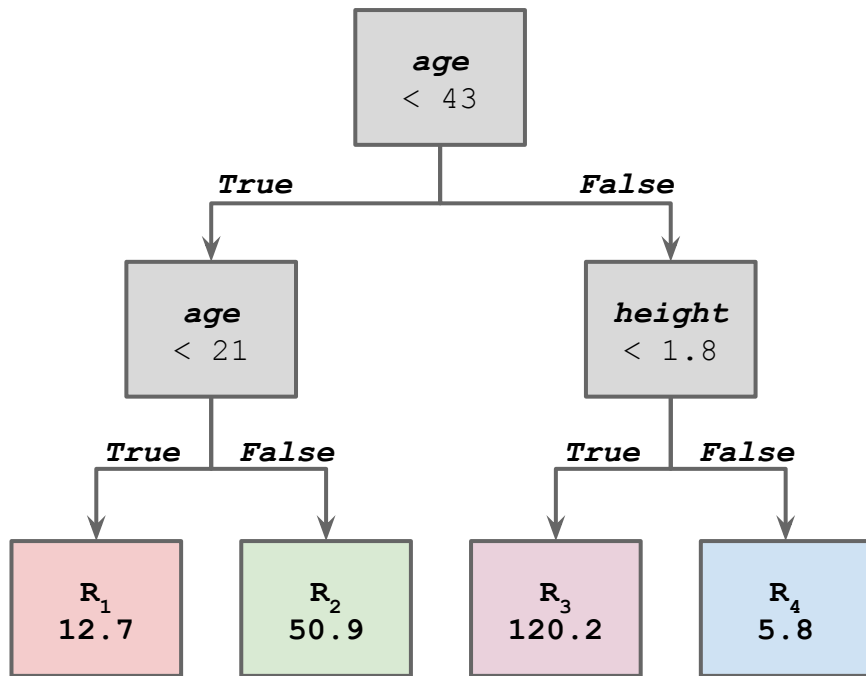
Assim como na Impureza de Gini e na Entropia, **quanto menor o RSS, maior a qualidade** de um conjunto de valores.



CONSTRUÇÃO DA ÁRVORE

O treinamento do modelo segue **o mesmo passo-a-passo** da construção da árvore de classificação, com a única diferença sendo o uso do **RSS** como função de homogeneidade.

OUTRA REPRESENTAÇÃO DOS *SPLITS*

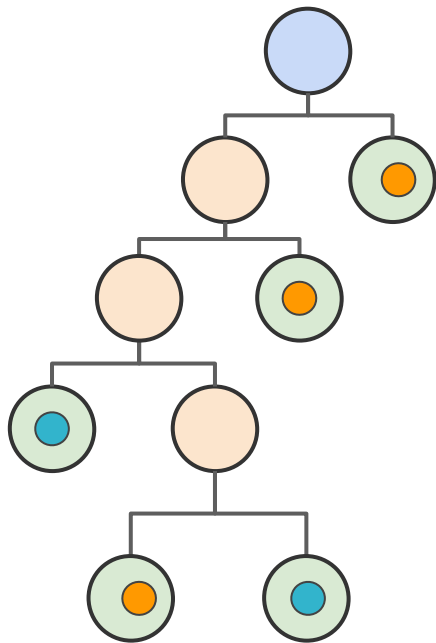


A vertical bar with a gradient from green at the top to blue at the bottom.

REGULARIZAÇÃO EM ÁRVORES

Técnicas para evitar o *overfit*

OVERFIT EM ÁRVORES



OVERFIT

Árvores de decisão também estão expostas ao **overfit**, falha de generalização em que o modelo aprende somente **características intrínsecas** aos dados de treinamento, não conseguindo ter bom desempenho em outros dados.

ÁRVORE INFINITA

Caso não haja uma **limitação explícita** no treinamento, uma árvore de decisão pode **criar um nó-folha por elemento** da massa de treino. Quando isso ocorre, há um **overfit** claro do modelo.

REGULARIZAÇÃO

Existem alguns **hiper-parâmetros** que podem ser usados para controlar a **altura da árvore** e a **quantidade de elementos** nas folhas. Também podem ser usados métodos de **poda de árvore**.

PARÂMETROS DE REGULARIZAÇÃO

Controle do *split* por arquitetura

max_depth: altura máxima permitida para a árvore

max_leaf_node: máximo de folhas permitido

min_samples_split: mínimo de elementos para permitir um *split*

min_samples_leaf: mínimo de elementos para criação de nó-folha

Controle do *split* por regras de decisão

max_features: máximo de features a serem observadas em cada *split*

min_impurity_decrease: permite *split* apenas se função *H* tiver um valor abaixo de um mínimo

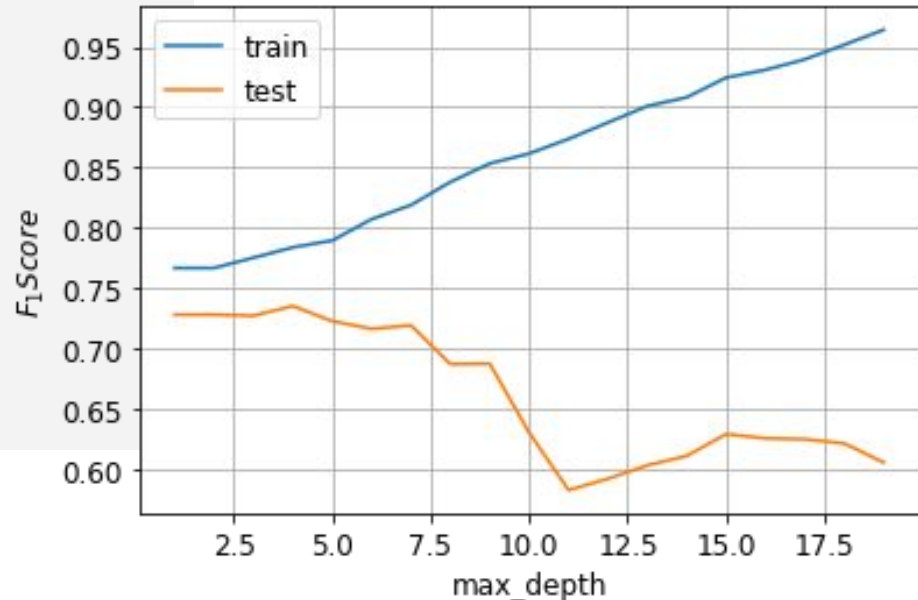
min_weight_fraction_leaf: controle para o caso de usar pesos (*weights*) nas folhas

Controle da poda de árvores

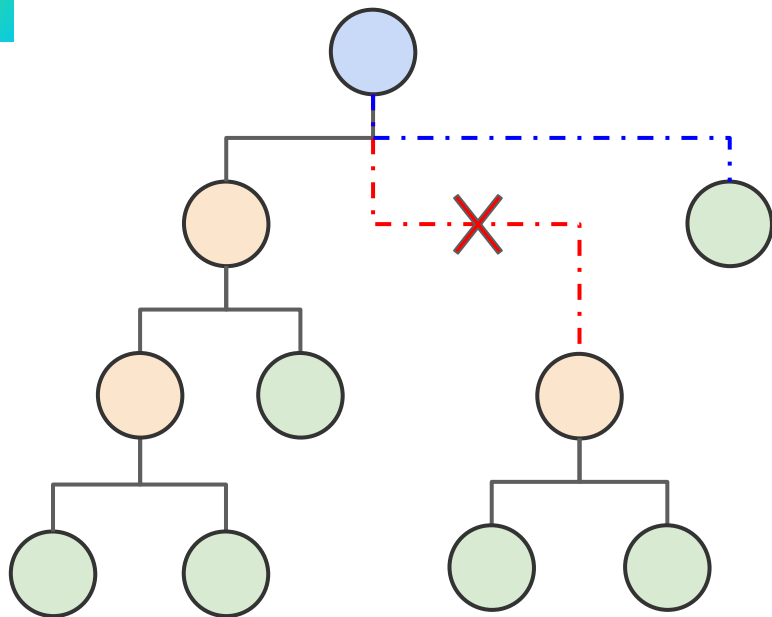
ccp_alpha: parâmetro de complexidade do método *Minimal Cost-Complexity Pruning*

```
f1_table = pd.DataFrame()
for max_depth in range(1, 20):
    model = DecisionTreeClassifier(max_depth=max_depth)
    model.fit(X_train, y_train)
    f1_tr = f1_score(y_train, model.predict(X_train))
    f1_te = f1_score(y_test, model.predict(X_test))
    f1_table = f1_table.append(
        pd.DataFrame(
            index=pd.Index(
                name='max_depth',
                data=[max_depth]
            ),
            columns=['train', 'test'],
            data=[[f1_tr, f1_te]]
        )
    )
f1_table.plot(grid=True)
plt.ylabel('$F_1$ Score$')
```

Efeito do parâmetro
max_depth no *overfit*



PODA EM ÁRVORES



RACIONAL

A poda de árvores serve para **reduzir a complexidade** final da árvore treinada. É uma técnica de **compressão de informação** muito utilizada em algoritmos de busca e em *machine learning*.

DIREÇÃO

A poda pode acontecer tanto da raiz para as folhas (**top-down**) quanto na direção contrária (**bottom-up**). Em ambos os casos, **nós de decisão** são avaliados em **termos de relevância** e, reprovados, são **substituídos por nós-folha**.

ALGORITMOS


O algoritmo **reduced error pruning** é o mais comum: **substitui aleatoriamente** sub-árvores pela **classe mais provável** e só reverte caso haja **queda de performance** relevante. A versão mais complexa é o **cost complexity pruning**: usa **estimativa de custo** para selecionar as sub-árvores.

T




RESUMO DA AULA

TAKEAWAY #1

A vertical bar with a gradient from green at the top to blue at the bottom.

A árvore de decisão é uma forma de representar os passos necessários para a resolução de um problema.


TAKEAWAY #2



Sempre treinamos a árvore de classificação com os conceitos de pureza e entropia.

O que fazemos é ver qual é a melhor divisão de dados em grupos que estão nos nós-folhas e que vão trazer o mínimo de ruído com o máximo de informação, tornando a árvore muito eficiente para a tomada de decisão.

TAKEAWAY #3



**Para as árvores de regressão,
nos focamos em média e
resíduos.**

Logo, queremos uma árvore de decisão que contenha as médias mais próximas dos elementos que compõem esta média.

TAKEAWAY #4

Por fim, quanto à regularização de árvores, vimos técnicas para evitar *overfit*, percebendo que não podemos deixar uma árvore livre para crescer para qualquer lado.

