



Data Science
Academy

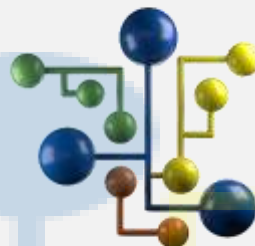
Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

Machine Learning



Data Science
Academy

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5



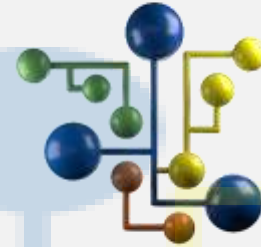
**Data Science
Academy**

Seja muito bem-vindo(a)!



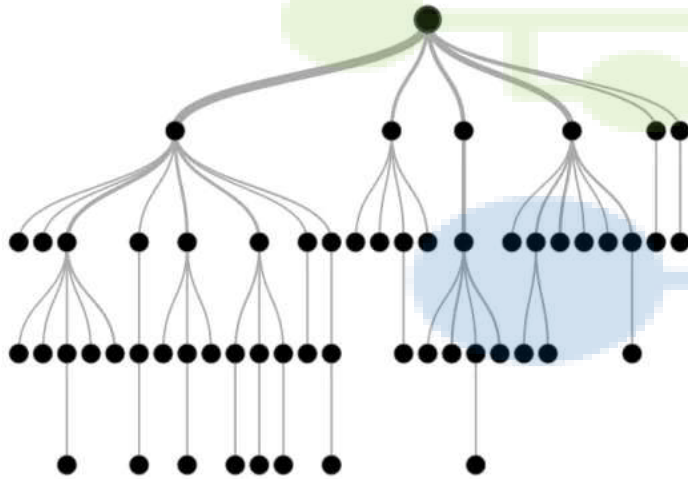
Data Science
Academy

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5



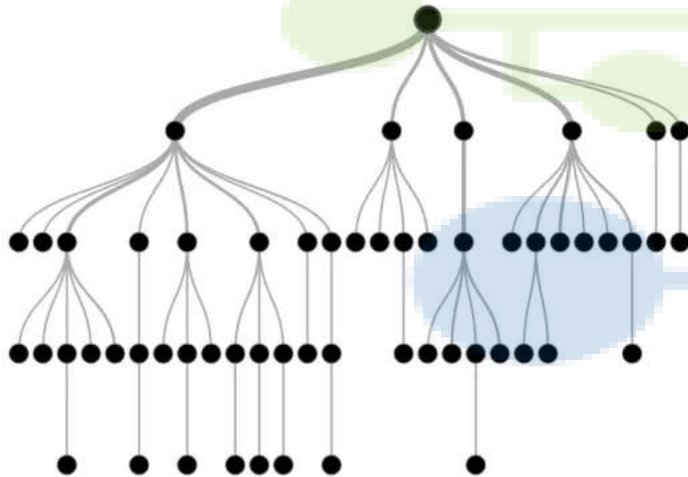
**Data Science
Academy**

Decision Tree, Random Forest e Métodos Ensemble

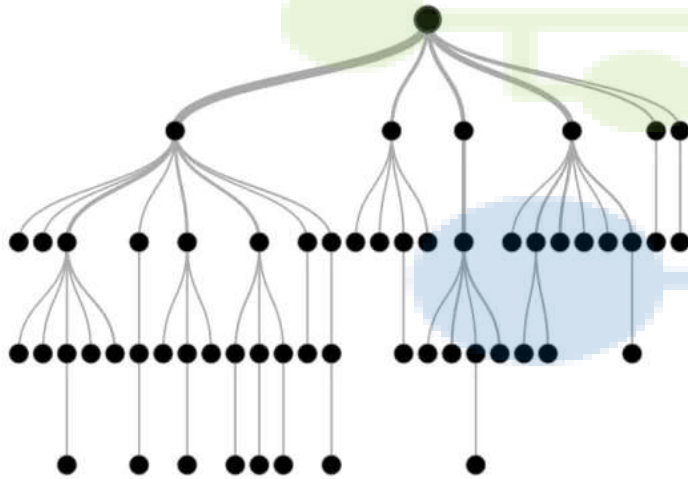


Este é um assunto bastante extenso, pois temos diversos algoritmos e diversas técnicas para trabalhar com árvores de decisão.

Por outro lado, esses algoritmos estão entre os mais poderosos em Machine Learning e são de fácil interpretação.

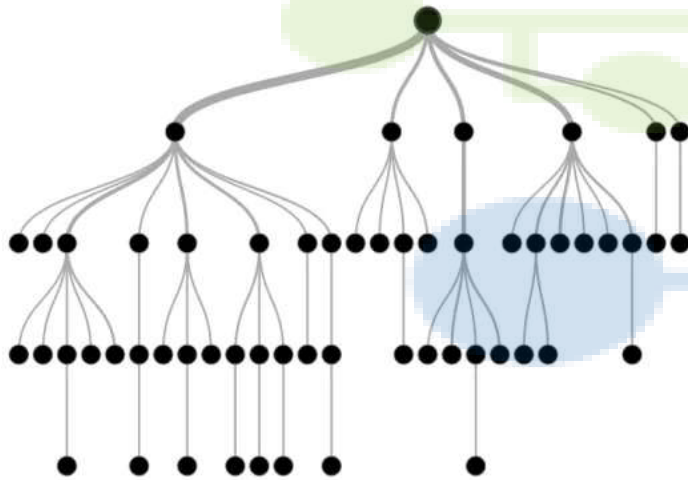


Vamos iniciar nossos estudos definindo o que são árvores de decisão e sua representação através de algoritmos de Machine Learning.



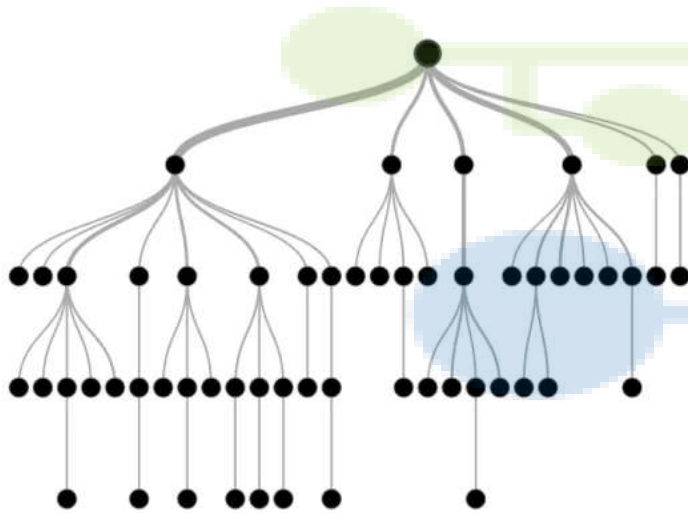
Como já conversamos nos capítulos anteriores, uma coisa é o modelo de aprendizagem e outra coisa é o algoritmo de aprendizagem.

Para os modelos de aprendizagem com árvores de decisão, estudaremos alguns algoritmos como o C4.5, C5.0, CART e o ID3.



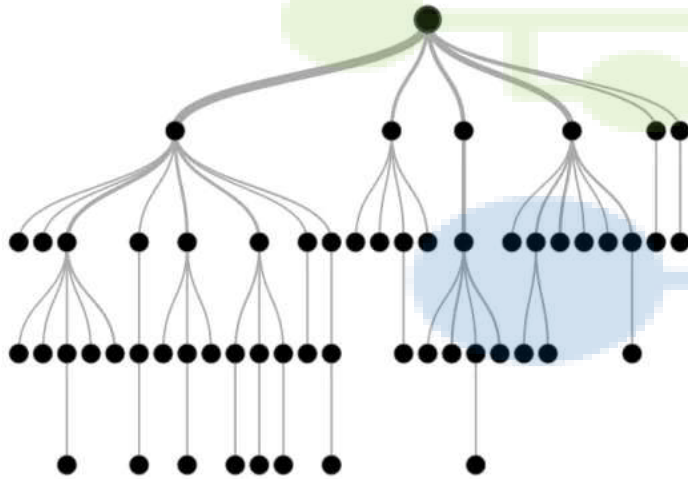
Existem alguns tipos especializados de árvores de decisão e estudaremos isso na sequência do capítulo.

E a principal especialização das árvores de decisão é o RandomForest, que nada mais é do que uma coleção de árvores de decisão. Estudaremos o RandomForest em detalhes.

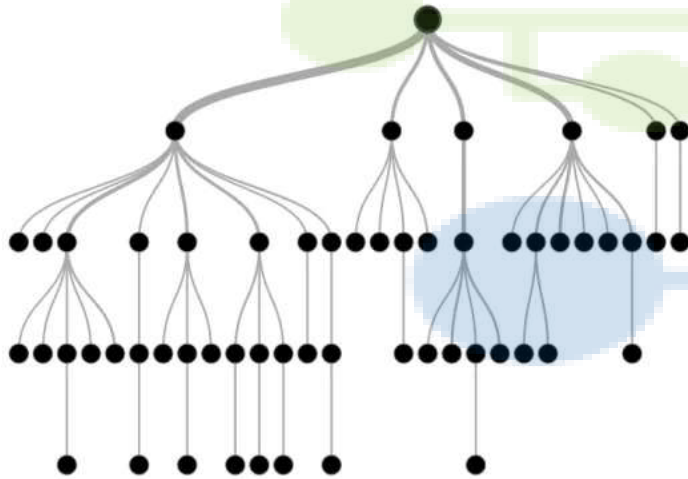


Podemos usar o RandomForest para seleção de atributos, ou seja, podemos usar árvores de decisão não apenas para modelos de ML em si, mas também para aplicar técnicas de feature selection a fim de preparar nosso dataset para outros algoritmos de ML.

Veremos os conceitos relacionados a seleção de atributos, tais como ganho de informação, entropia e índice Gini.



E vamos claro criar modelos e fazer previsões, estudar os parâmetros e os detalhes de pré-processamento das árvores de decisão e como interpretar os resultados dos modelos preditivos.

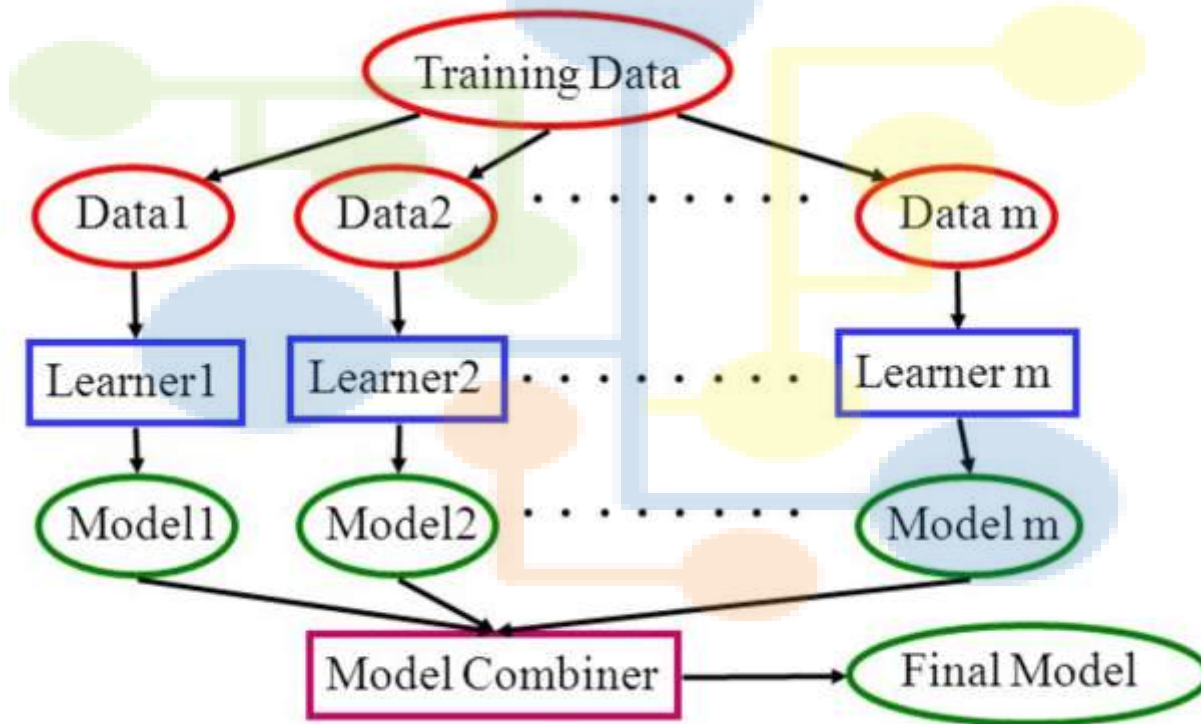


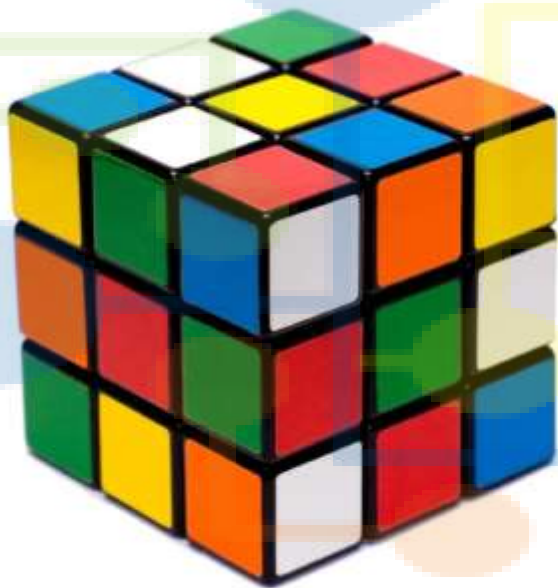
E faremos ainda o pruning, que em português seria algo como “podar a árvore”.

Ao criarmos árvores de decisão, podemos ter árvores com muitos “galhos e folhas” e em algum momento teremos que parar a construção da árvore ou fazer ajustes reduzindo o número de pontos de decisão no modelo preditivo. Veremos como aplicar esta técnica.



Métodos Ensemble

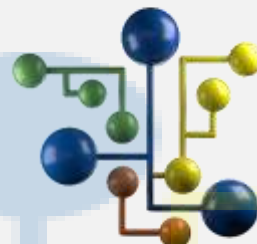






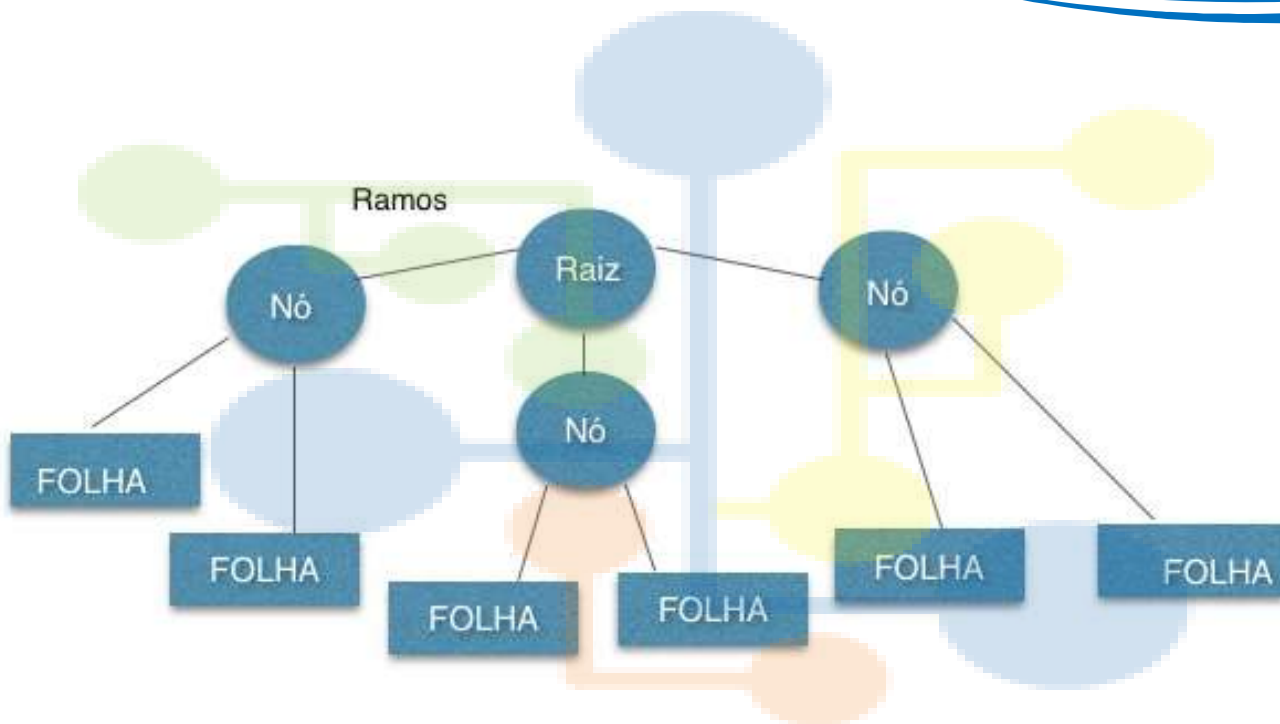
Data Science
Academy

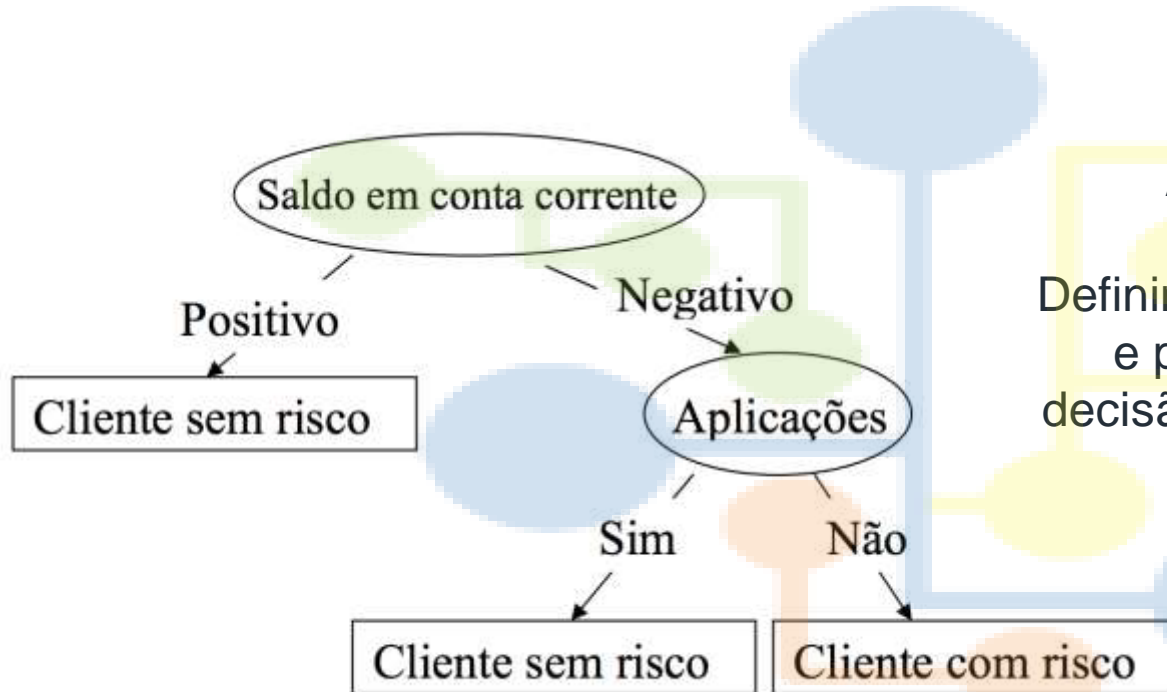
Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5



**Data Science
Academy**

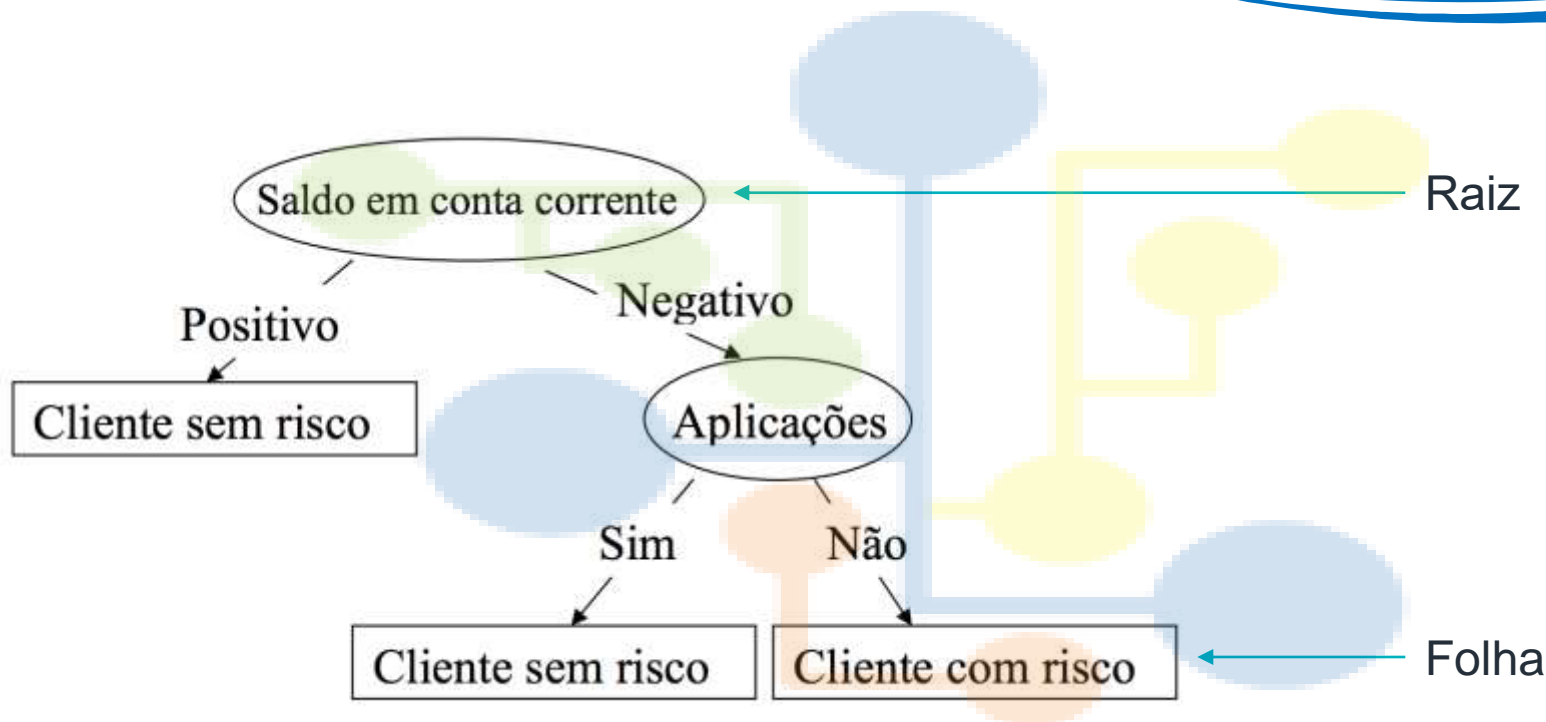
Árvores de Decisão

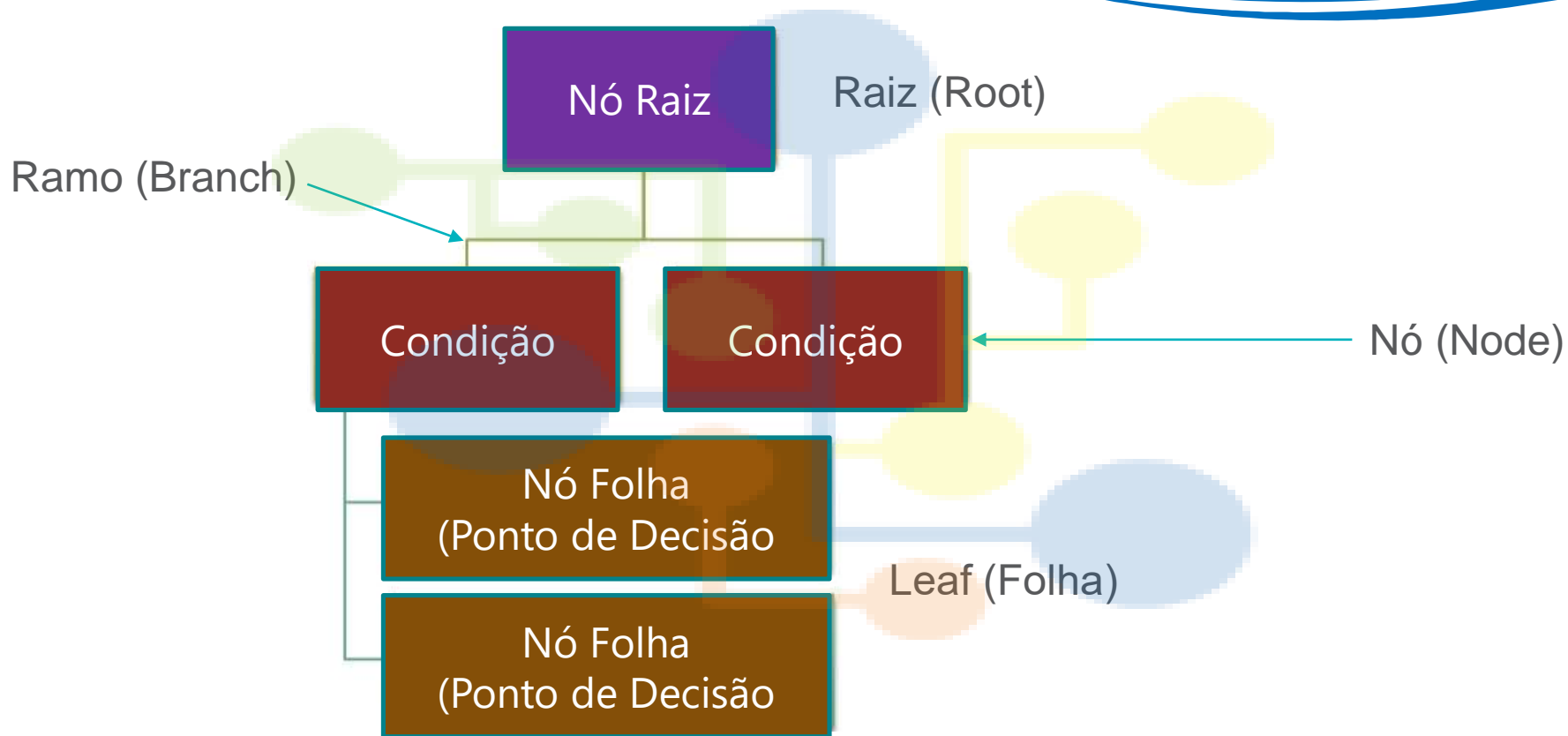


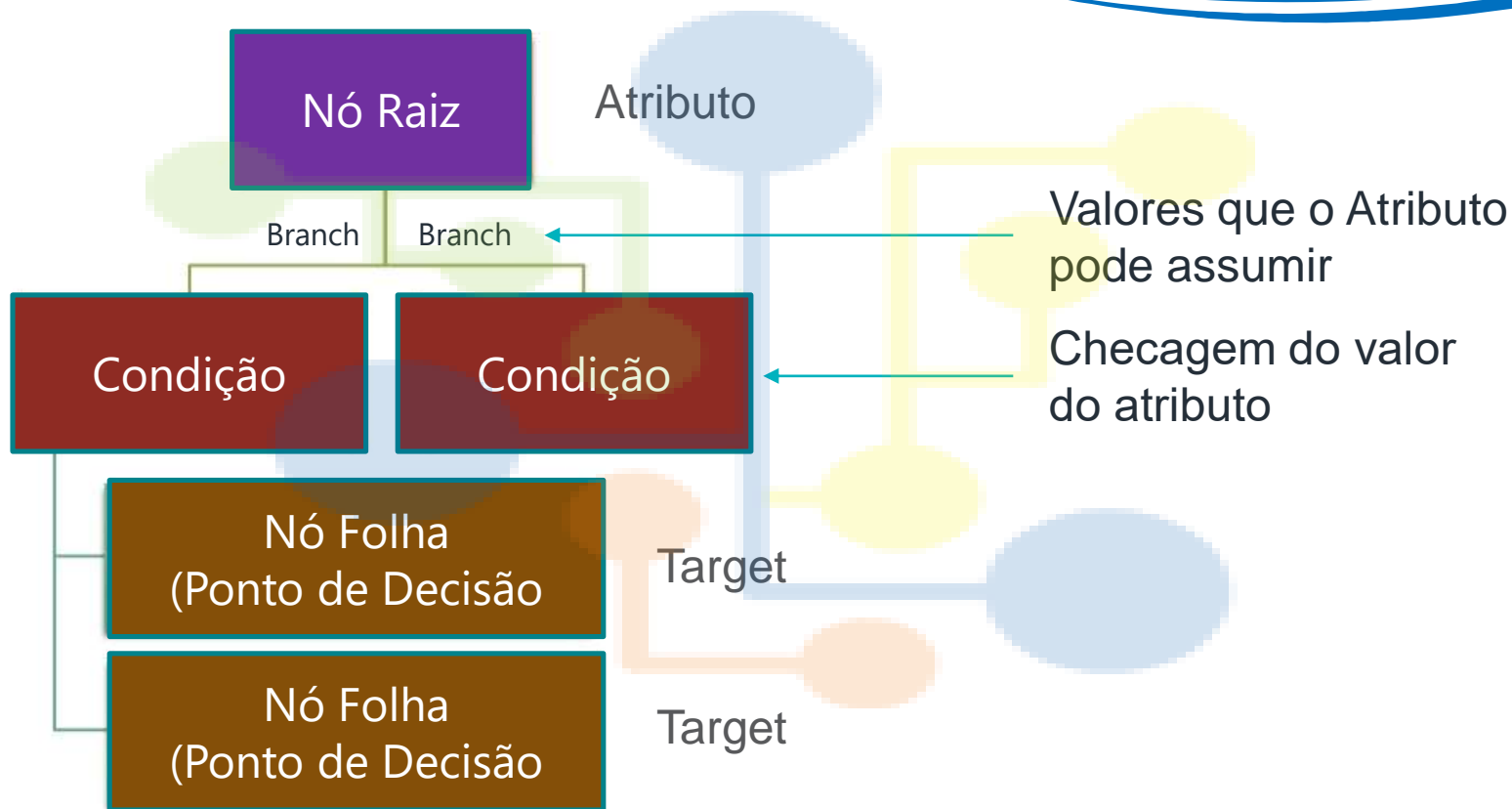


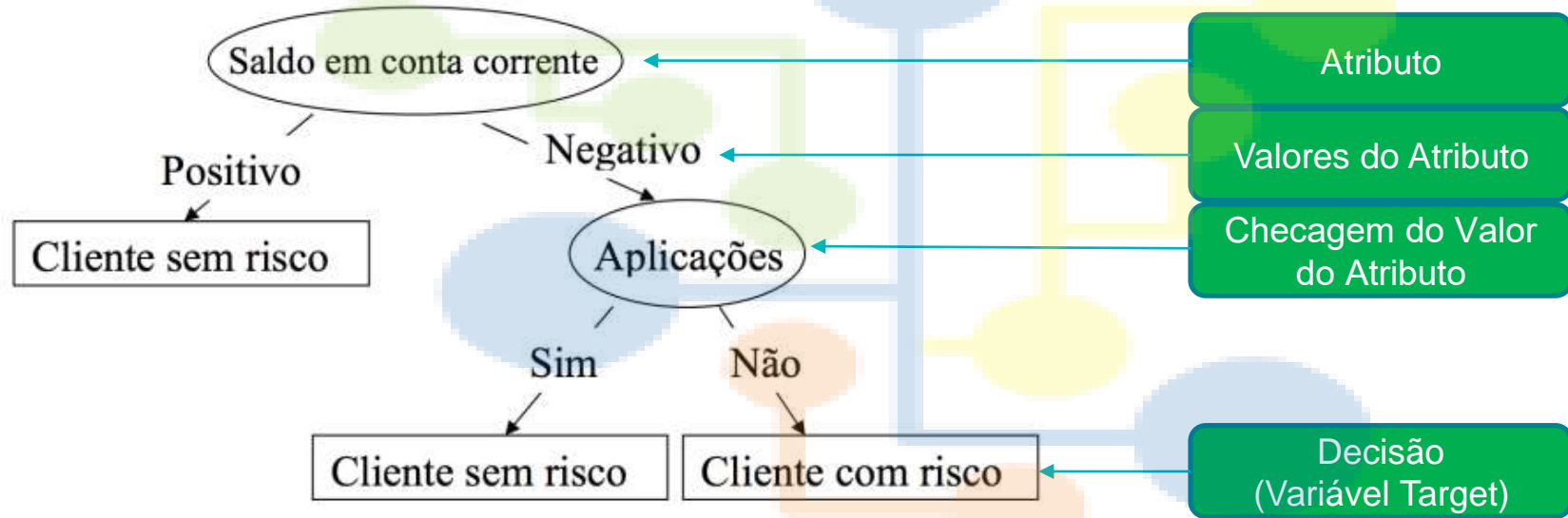
Árvores de Decisão

Definimos um conjunto de regras e para cada regra há uma decisão que precisa ser tomada.











Árvores de Decisão podem ser usadas para problemas de:

Classificação

**Árvore de
Classificação**

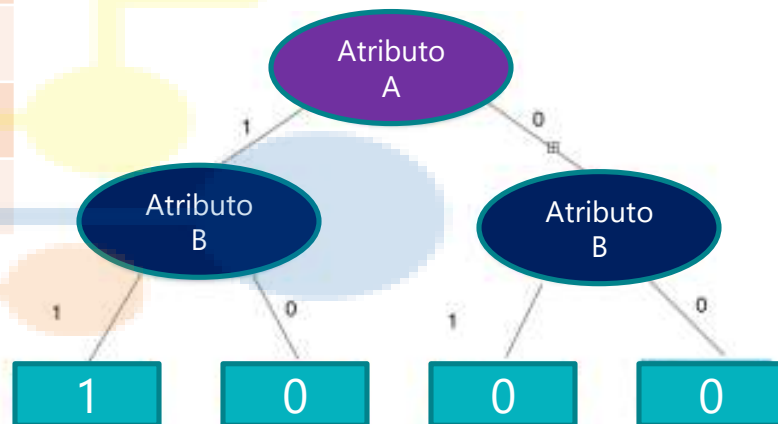
Regressão

**Árvore de
Regressão**



Considerações na Construção de Árvores de Decisão

Atributo A	Atributo B	Saída
0	0	0
0	1	0
1	0	0
1	1	1



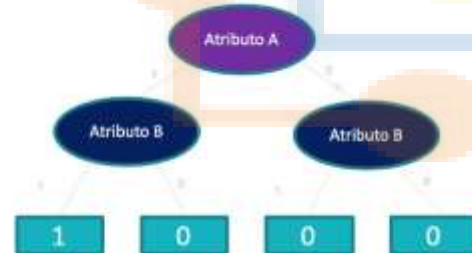


Considerações na Construção de Árvores de Decisão

Qual atributo deve ser usado para iniciar a árvore?

Qual deve ser o atributo seguinte?

Quando parar de construir ramos na árvore (para evitar overfitting)?





Considerações na Construção de Árvores de Decisão

Qual atributo deve ser usado para iniciar a árvore?

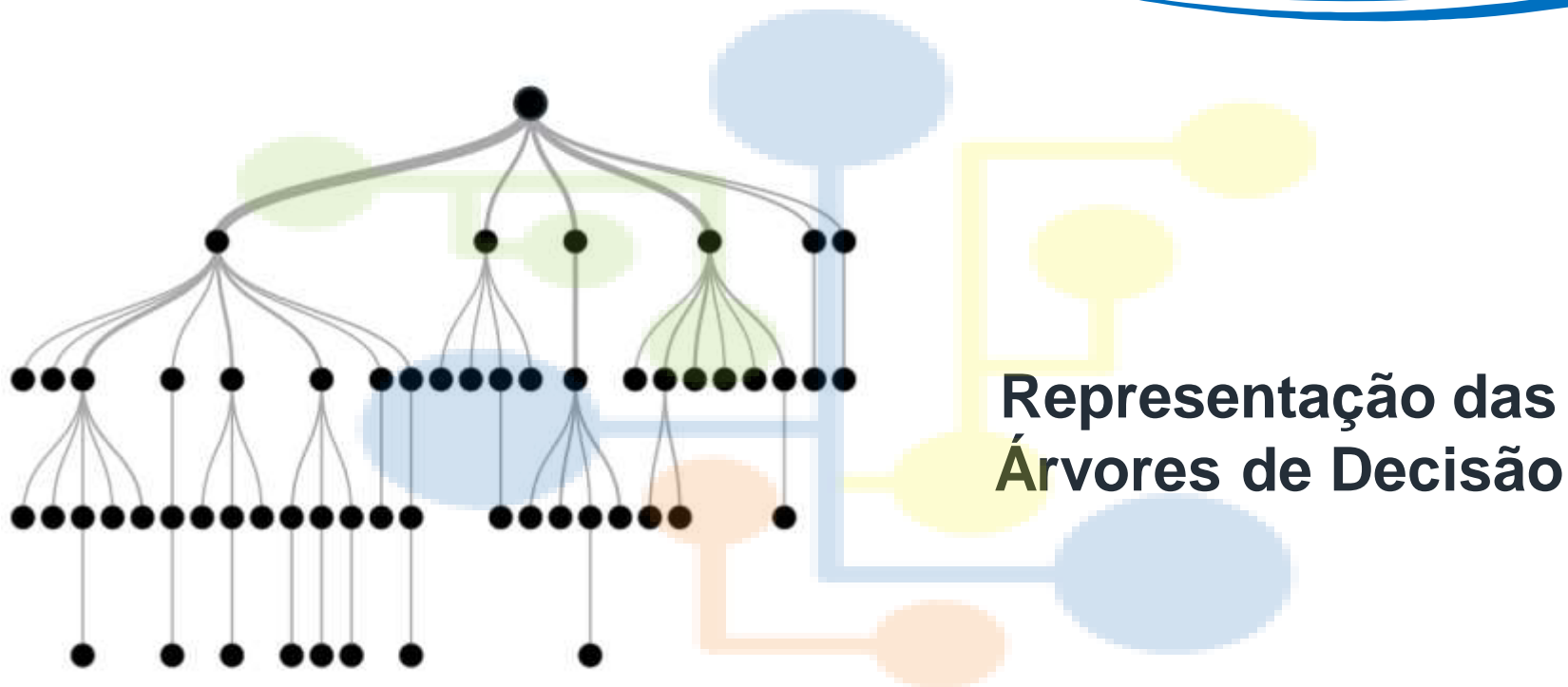
Qual deve ser o atributo seguinte?

Quando parar de construir ramos na árvore (para evitar overfitting)?

Ganho de Informação e
Entropia

Índice de Gini
(Gini Index)

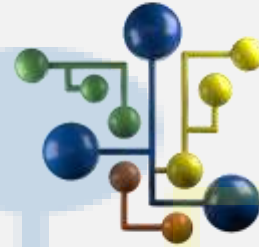
Taxa de Ganho
(Gain Ratio)





Data Science
Academy

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5

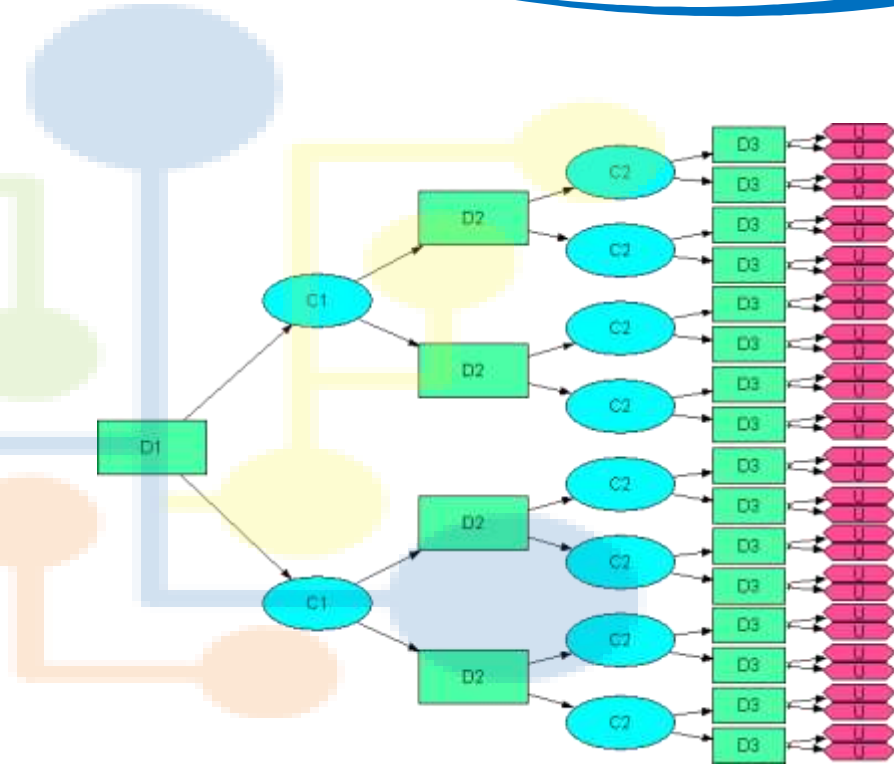


**Data Science
Academy**

Ganho de Informação, Entropia, Índice Gini e Pruning



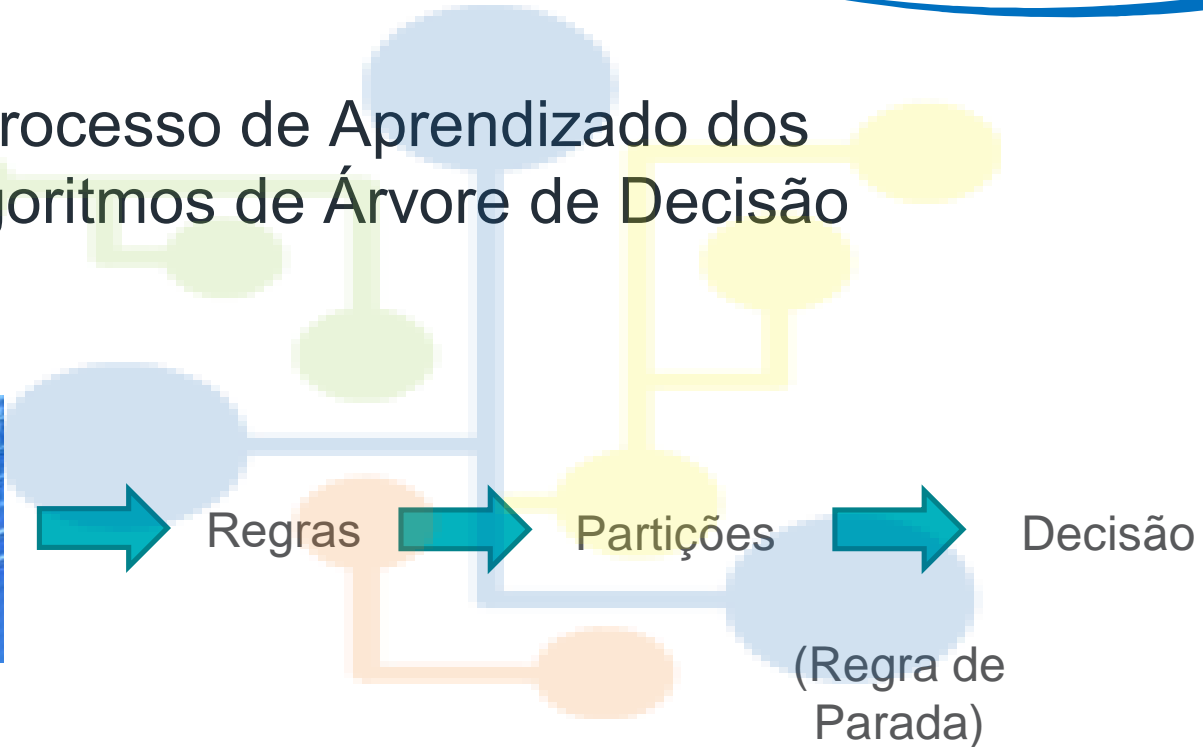
As árvores de decisão têm desfrutado de muita popularidade por causa de seu algoritmo intuitivo. Sua saída é facilmente traduzida em regras e, portanto, é bastante compreensível pelos seres humanos (diferente de modelos como SVM e Redes Neurais, consideradas caixas pretas).

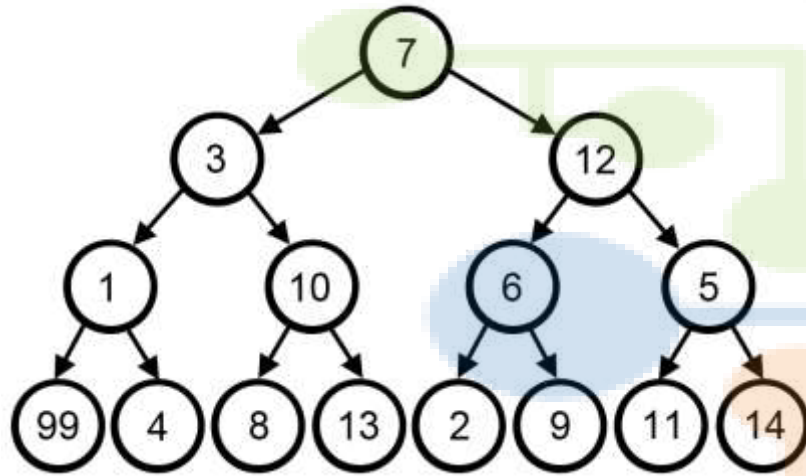


Processo de Aprendizado dos Algoritmos de Árvore de Decisão



Dados





Greedy Search (Busca Gananciosa ou Gulosa)

O algoritmo procura maximizar o passo atual sem olhar para o passo seguinte, a fim de alcançar uma otimização global.



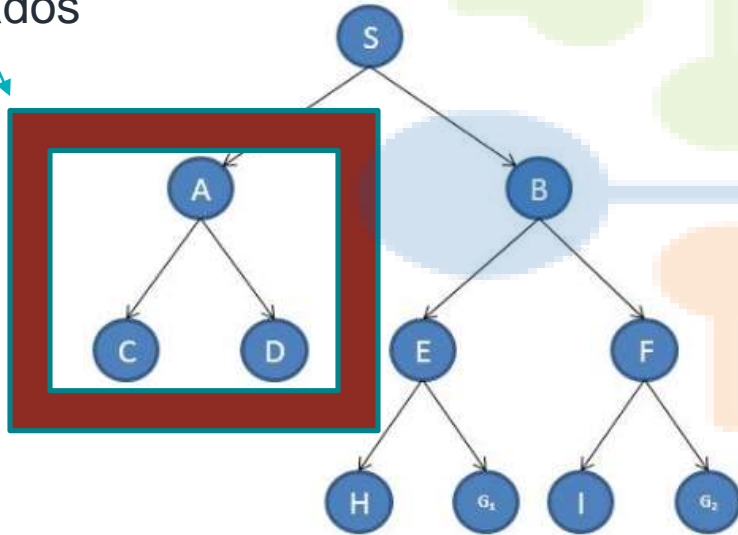
**Mais detalhes sobre algoritmos gulosos
no curso de Introdução à Lógica de
Programação disponível para os alunos
das Formações DSA.**



Greedy Search utiliza uma heurística estimada $h(n)$

Partição de Dados

S → Estado inicial
G1, G2 → Objetivo



Node	$h(n)$
A	11
B	5
C	9
D	8
E	4
F	2
H	7
I	3



Índice Gini

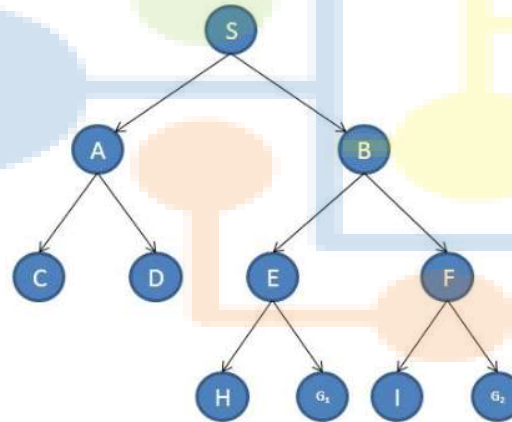
Ganho de
Informação

Redução de
Variância

Ross Quinlan \rightarrow (ID3) \rightarrow C4.5 \rightarrow C5.0



Como definir o nó raiz e como realizar a divisão do conjunto de dados?



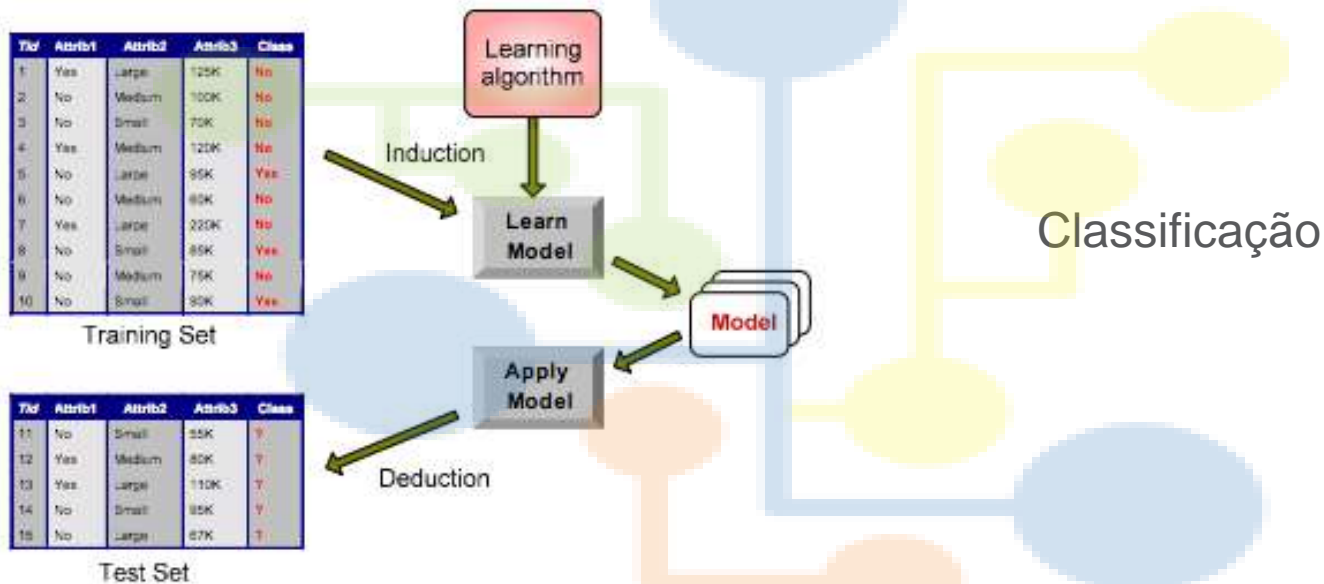


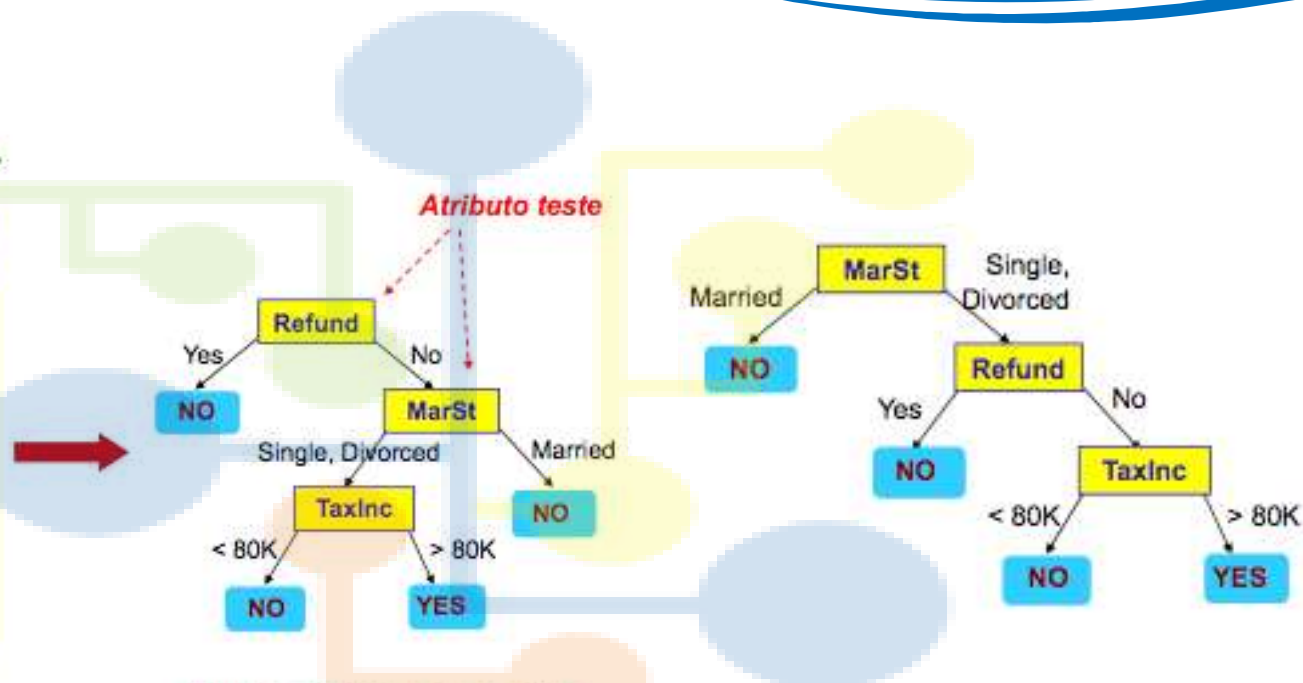


Diagram illustrating the data types for the training dataset:

- Refund**: categorical
- Marital Status**: categorical
- Taxable Income**: continuous
- Cheat**: class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Dados de
treinamento



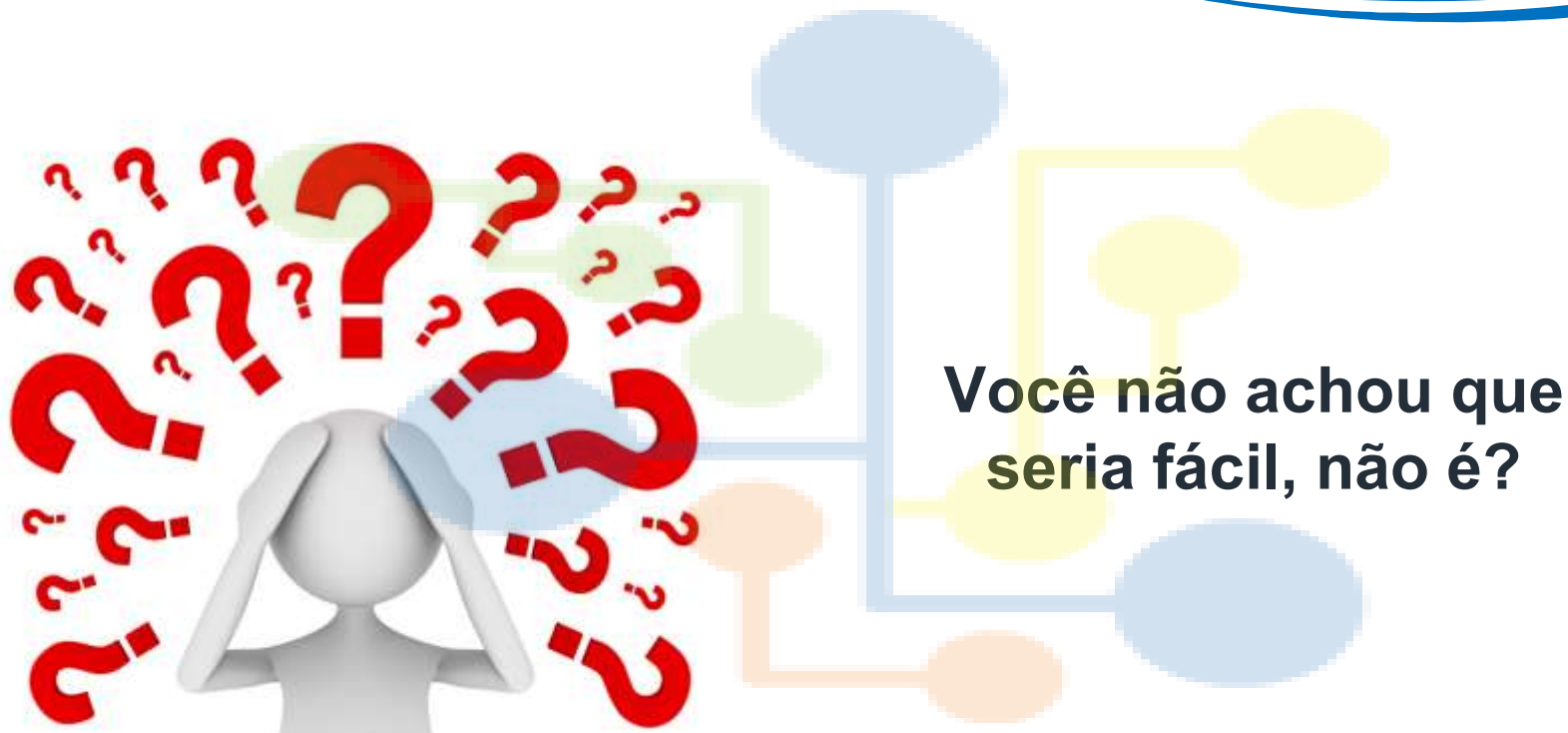
Modelo: árvore de decisão

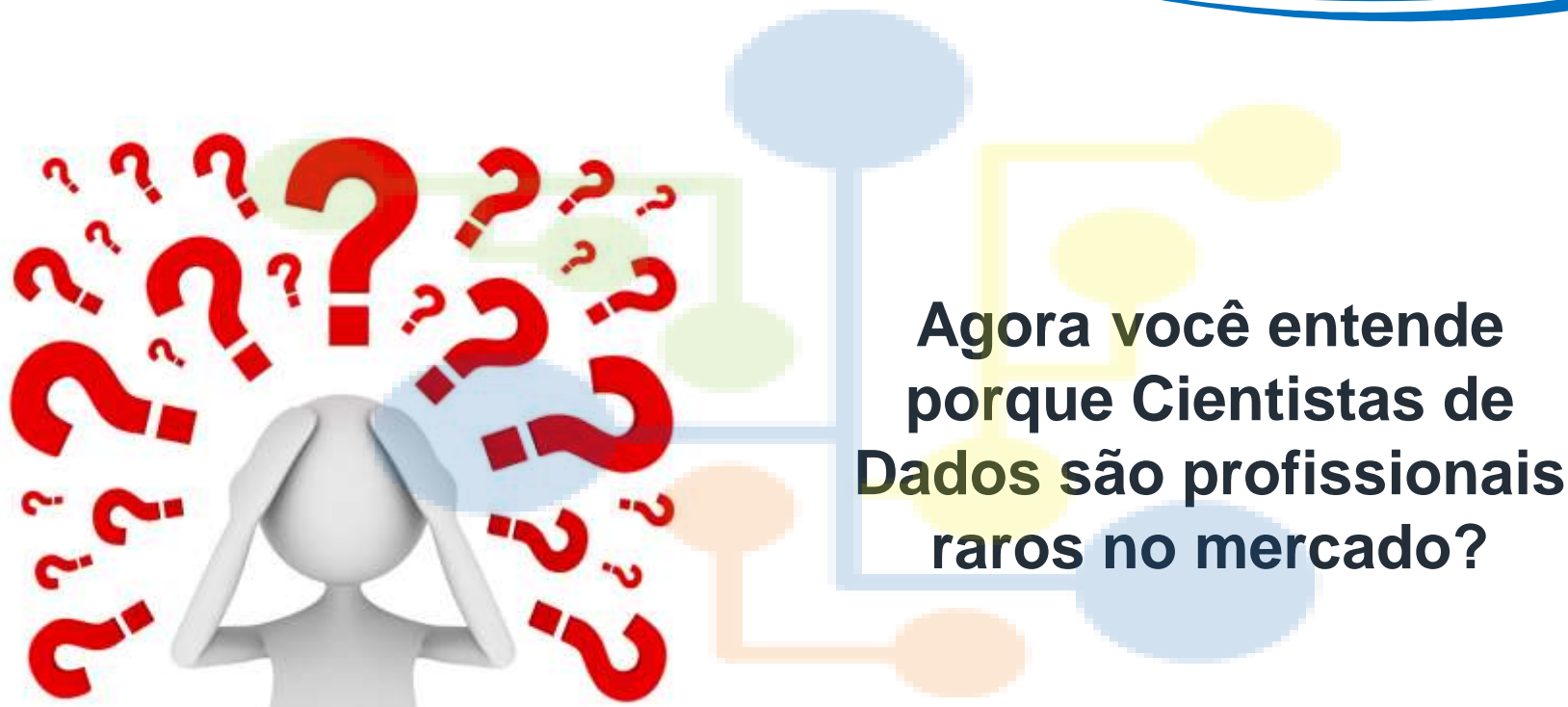
Pode haver mais de um árvore para o
mesmo conjunto de dados

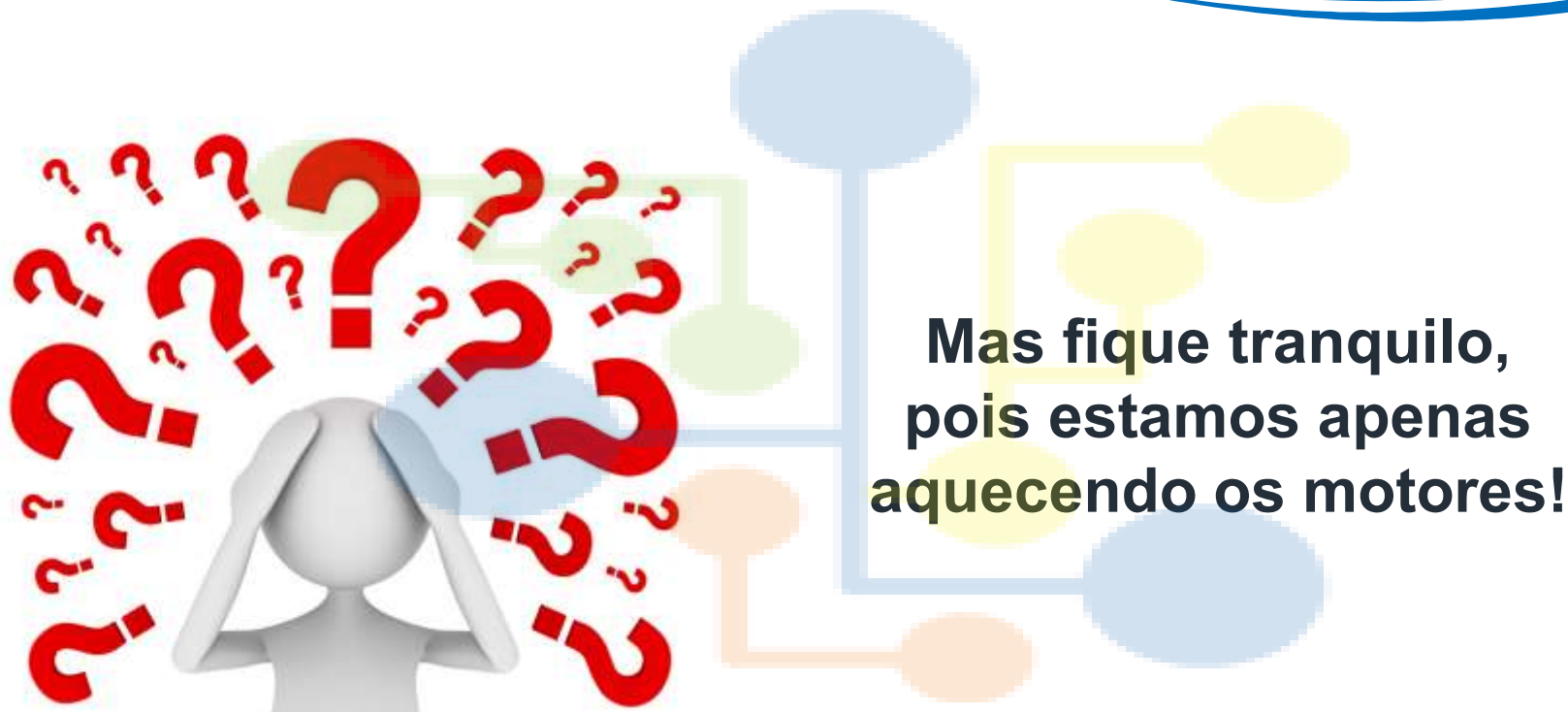


Como definir o nó raiz e como realizar a divisão do conjunto de dados?

- Estratégia Gulosa (Greedy Selection)
- Divisão baseada em atributos nominais
 - Divisão Binária
 - Divisão Múltipla
- Divisão baseada em atributos contínuos
 - Decisão Binária
 - Discretização
 - Estática
 - Dinâmica







**Mas fique tranquilo,
pois estamos apenas
aquecendo os motores!**



Como definir o nó raiz e como realizar a divisão do conjunto de dados?

Estratégia Gulosa (Greedy Selection)

Necessita da medida da “impureza” do nó

C0: 5
C1: 5

Não-homogênea,
Alto grau de impureza

C0: 9
C1: 1

Homogêneo,
baixo grau de impureza



Como definir o nó raiz e como realizar a divisão do conjunto de dados?

Estratégia Gulosa (Greedy Selection)

Necessita da medida da “impureza” do nó

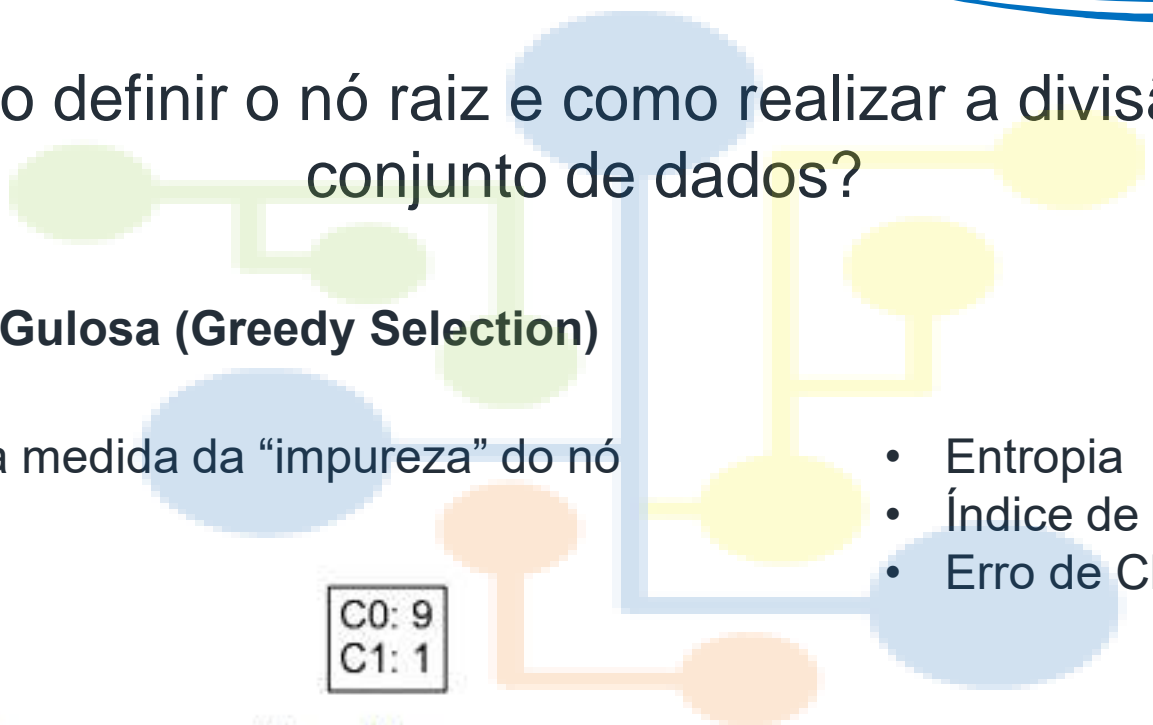
- Entropia
- Índice de Gini
- Erro de Classificação

C0: 5
C1: 5

Não-homogênea,
Alto grau de impureza

C0: 9
C1: 1

Homogêneo,
baixo grau de impureza





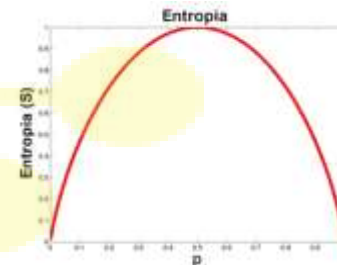
Entropia é a medida da incerteza nos dados

Ganho de Informação é a redução da Entropia



Entropia

$$Entropy = \sum -p_i \log_2 p_i$$



Entropia máxima considerando duas classes com a mesma probabilidade (distribuição 50/50):

$$Entropy = -0.5 * \log_2(0.5) - 0.5 * \log_2(0.5) = 1.0$$

Entropia considerando duas classes com distribuição 40/60:

$$Entropy = -0.4 * \log_2(0.4) - 0.6 * \log_2(0.6) = 0.97$$



Nos algoritmos ID3, C4.5 e C5.0, o nó raiz é escolhido com base em quanto do total da Entropia é reduzido, se aquele nó é escolhido

Isso é chamado de Ganho de Informação!



Ganho de Informação = Entropia do sistema antes da divisão – Entropia do sistema após a divisão





Ganho de Informação = Entropia do sistema antes da divisão – Entropia do sistema após a divisão

$$E = - \sum_{i=1}^m p_i \log_2(p_i)$$

$$E_A = \sum_{i=1}^v \frac{D_i}{D} E(D_i)$$



Esta metodologia (Entropia) é aplicada para computar o ganho de informação para todos os atributos. É escolhido o atributo com o mais alto ganho de informação. Isso é testado para cada nó a fim de escolher o melhor nó.



Índice de Gini

O Índice de Gini é usado para medir a probabilidade de dois itens aleatórios pertencerem à mesma classe.

A medida de Gini de um nó é a soma dos quadrados das proporções das classes.



Índice de Gini

O Índice de Gini diz: se selecionarmos dois itens de uma população aleatoriamente, então eles devem ser da mesma classe e a probabilidade para isto é 1 se a população é pura.

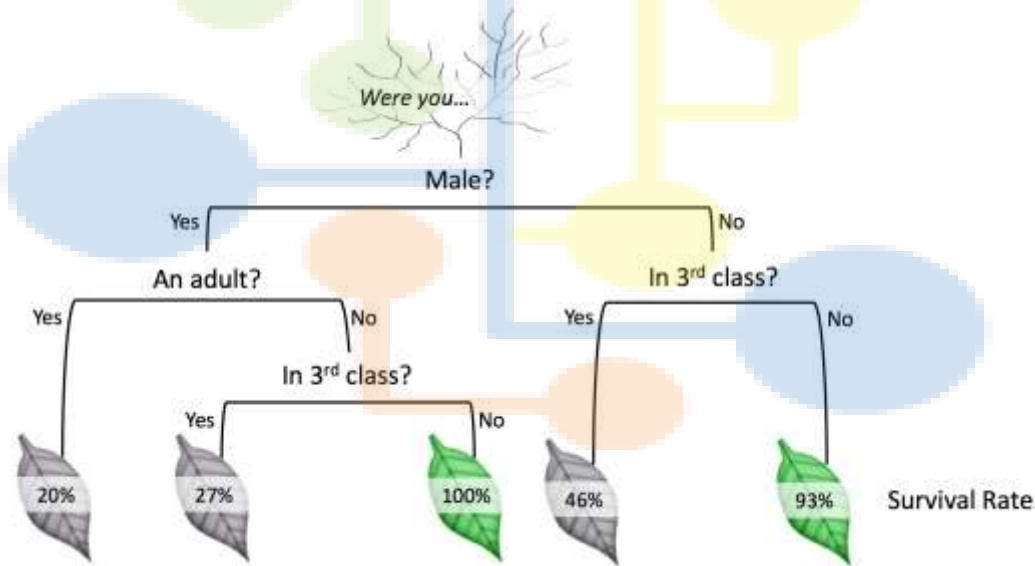


Índice de Gini

O Índice de Gini é usado como regra de parada para construção de uma árvore de decisão.



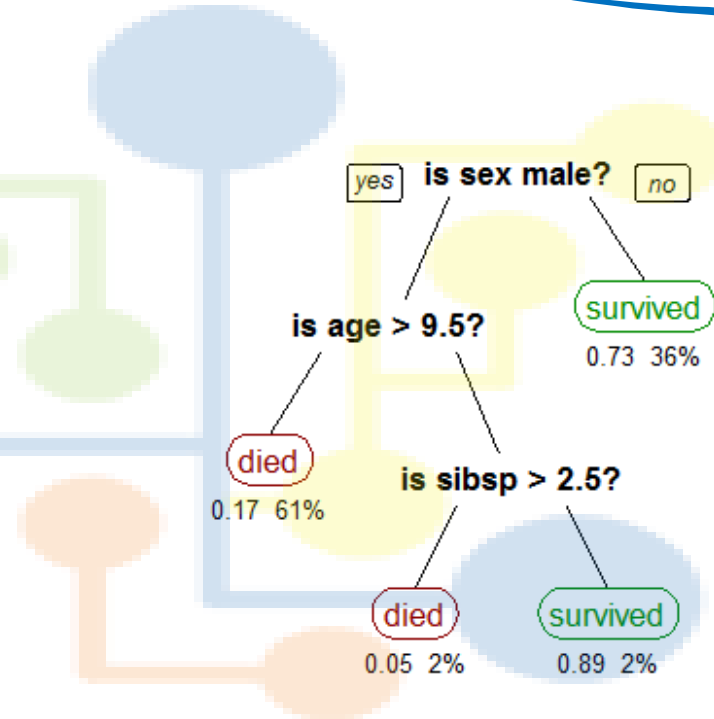
O que são as regras de parada (Stopping Rules)?





Regras de Parada

- Índice Gini
- Qui-quadrado
- Ganho de Informação
- Redução de Variância





Pruning
Poda da Árvore





Pruning

- A árvore de decisão é concluída antes que uma classificação perfeita dos dados de treinamento seja alcançada.
- Ocorre o excesso de ajuste nos dados gerando um modelo e, em seguida, a árvore é podada (Pruning) para se tornar generalizável.



E como definir o tamanho correto da árvore?

Usar um conjunto
de validação

Usar métodos
probabilísticos



O classificador de árvore de decisão do Scikit-Learn não suporta atualmente o Pruning. Pacotes avançados como o XGBoost adotaram a poda de árvores em sua implementação. Mas a biblioteca rpart em R, fornece uma função para Pruning.

Viu por que é importante conhecer mais de uma ferramenta?



Data Science
Academy

Data Science Academy thaylathais1@gmail.com 611665e6e32fc3dbb86d0db5



**Data Science
Academy**

Algoritmo ID3

An abstract background diagram consisting of several colored circles (blue, green, yellow, orange) connected by lines, forming a network-like structure. The text "O que são heurísticas?" is overlaid on this diagram.

O que são heurísticas?



Começa com todos os exemplos de treino

Escolhe o teste (atributo) que melhor divide os exemplos, ou seja agrupa exemplos da mesma classe ou exemplos semelhantes

Para o atributo escolhido, é criado um nó filho para cada valor possível do atributo

Transporta os exemplos para cada filho considerando o valor do filho

Repete o procedimento para cada filho não "puro".



E como o algoritmo sabe o melhor atributo a escolher?

Através do Ganho de Informação e Entropia!!



A heurística ID3

A decorative background diagram consisting of a central vertical blue line with several horizontal and diagonal branches. At the end of these branches are circles of various colors: blue, green, yellow, and orange. The circles are of different sizes and are arranged in a way that suggests a hierarchical or branching structure, similar to a decision tree.



Espaço de Hipóteses do ID3

A decorative background diagram consisting of a central vertical blue line with several horizontal and vertical branches. At the end of these branches are circular nodes in various colors: blue, green, yellow, and orange. The nodes are connected by lines of the same color as the nodes they connect, creating a complex, branching structure that resembles a decision tree or a network graph.



A faint, stylized decision tree diagram serves as a background. It features a central vertical line with several horizontal branches. The branches are color-coded: green for the top-left, blue for the top-right, orange for the bottom-left, and yellow for the bottom-right. Each branch terminates in a circular node of the same color. The nodes are arranged in a hierarchical structure, with the top-left node being the largest and most prominent.

ID3 (Iterative Dichotomizer 3)

C4.5

C5.0

CART
(Classification and Regression Trees)



Continue Trilhando uma Excelente Jornada de Aprendizagem!

Muito Obrigado!