

**ÁRVORES, REDES E ENSEMBLE
MODELS I**

João F. Serrajordia R. de Mello

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.

Proibida a reprodução, total ou parcial, sem autorização. Lei nº 9610/98

Apresentação

João Fernando Serrajordia Rocha de Mello – (Juka)

Trajetória profissional

Modelagem de crédito em grandes bancos

Telecom

Desenvolvimento de modelos / Validação de modelos

Docência em ciência de dados

Consultoria em ciência de dados

Outsourcing executivo

Acadêmico



BACHAREL EM ESTATÍSTICA

MESTRE EM ESTATÍSTICA



O que vamos aprender?

- Árvores de decisão
- Ensemble
 - Bagging (Random Forest)
 - Boosting (Gradient Boosting)
 - Princípios de validação cruzada
- Introdução a Redes Neurais



Agenda de hoje

- Conceituação
- Árvores de decisão
 - Aplicações
 - O que é, como roda etc...
 - Detalhes operacionais (base etc)
 - Algoritmo
 - Como avaliar?
 - Problemas comuns





Vamos discutir...

O que é aprendizado de máquina
(o machine learning)?



O que é Machine Learning?

- É um ramo da Inteligência Artificial?
- Calma... O que é Inteligência Artificial afinal de contas?

Definição mais ‘mundana’



Senso comum:

Inteligência artificial é a área que estuda a implementação de atividades realizadas por máquinas, que anteriormente eram realizadas por humanos.



Dicionário Oxford:

“the theory and development of computer systems able to perform tasks that no [humans] can do or that can do in a way that is significantly faster or more accurate than a human can”
“a teoria e o desenvolvimento de sistemas computacionais capazes de executar tarefas que normalmente requerem inteligência humana, como percepção visual, reconhecimento de fala, tomada de decisão e tradução entre idiomas.”



Definição do MIT

- “A inteligência artificial é a capacidade dos computadores de imitar funções cognitivas humanas, como aprendizado e solução de problemas. Por meio da IA, um sistema de computador usa matemática e lógica para simular o raciocínio que as pessoas usam para aprender com novas informações e tomar decisões.”

<https://mitsloan.mit.edu/>

Definição acadêmica

- “We define AI as the study of agents that receive percepts from the environment and perform actions.”
Artificial Intelligence: a modern approach, de Stuart Russell e Peter Norvig

“Definimos IA como o estudo de agentes que recebem percepções do ambiente e realizam ações.”
<http://aima.cs.berkeley.edu/>





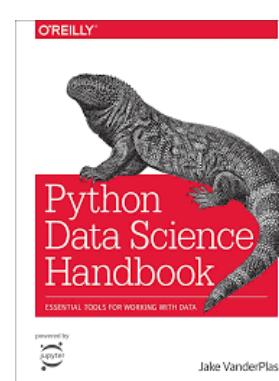
Vamos discutir...

E o que é aprendizado de máquina...?

Aprendizado de máquina

- Jake Van Der Plas:
 - Ajuda mais pensar no machine learning como uma forma de construir modelos sobre dados.
 - Envolve a construção de modelos matemáticos para entender dados.
 - O “aprendizado” entra na possibilidade de se obter parâmetros reguláveis que podem ser adequados aos dados observados.

<https://jakevdp.github.io/PythonDataScienceHandbook/05.01-what-is-machine-learning.html>



Você vai precisar de...



Preparativos

- Abrir o R
- Importar as bibliotecas
- Planilha eletrônica



Árvores de decisão



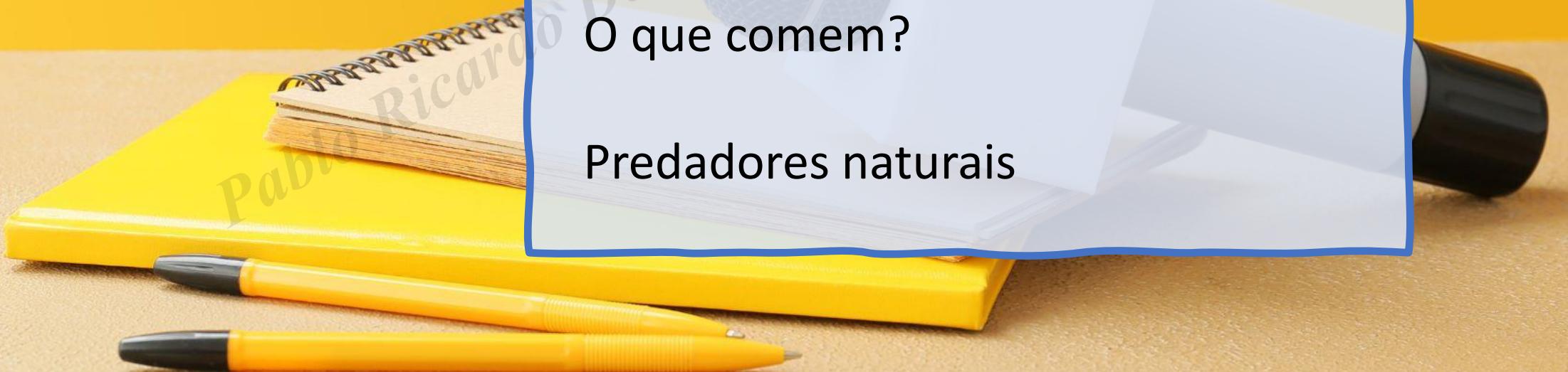
Árvores de decisão:

Onde vivem?

O que são?

O que comem?

Predadores naturais



Aplicações



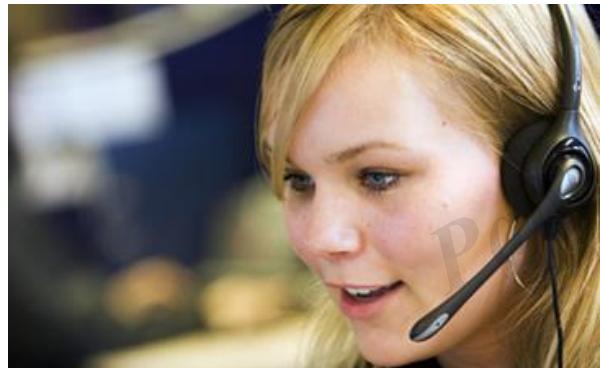
Qual a eficácia de uma vacina?



O cliente vai pagar o empréstimo?



Quanto de petróleo tem no poço?



O cliente vai comprar meu produto?

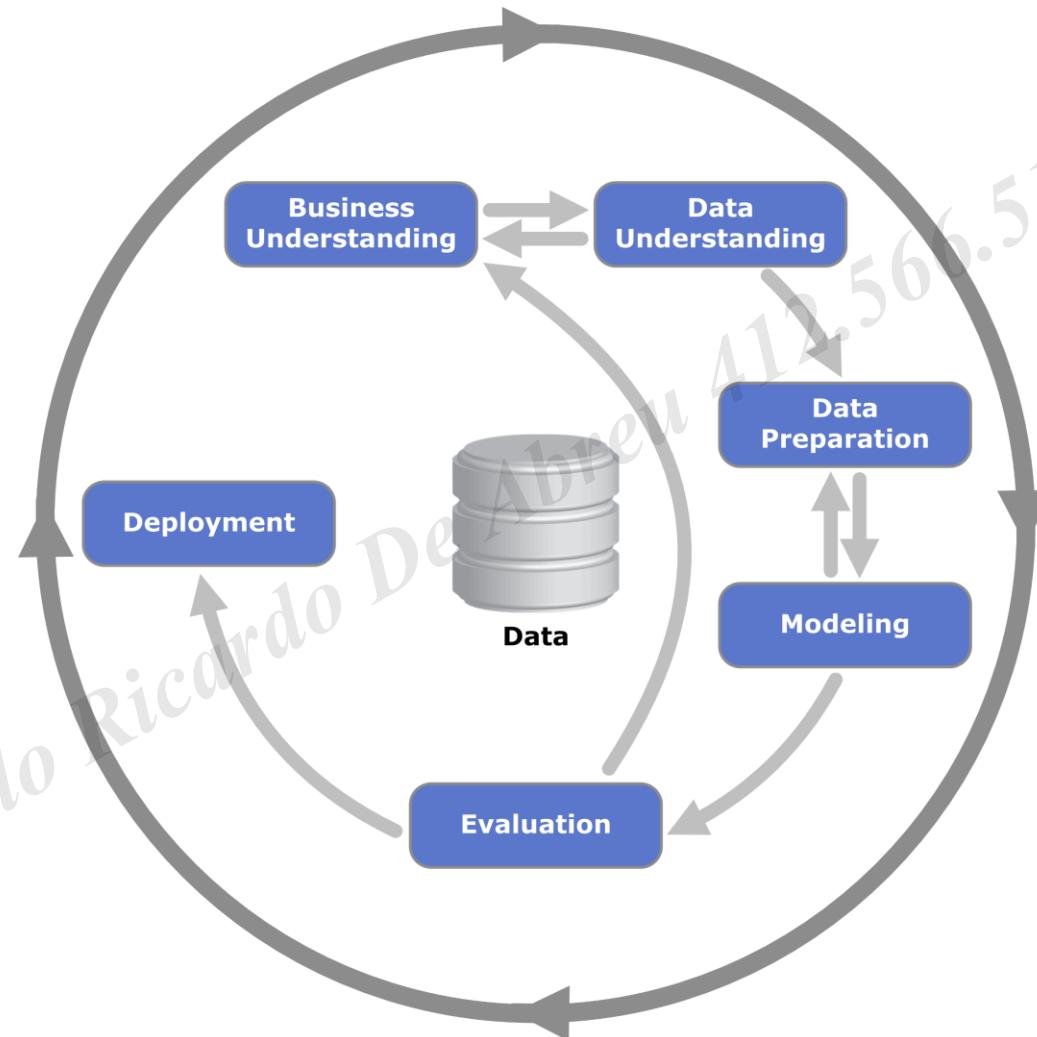


O que a pessoa está fazendo?



Quão ecológico esse veículo é?

CRISP-DM



Fonte: <https://www.the-modeling-agency.com/crisp-dm.pdf>

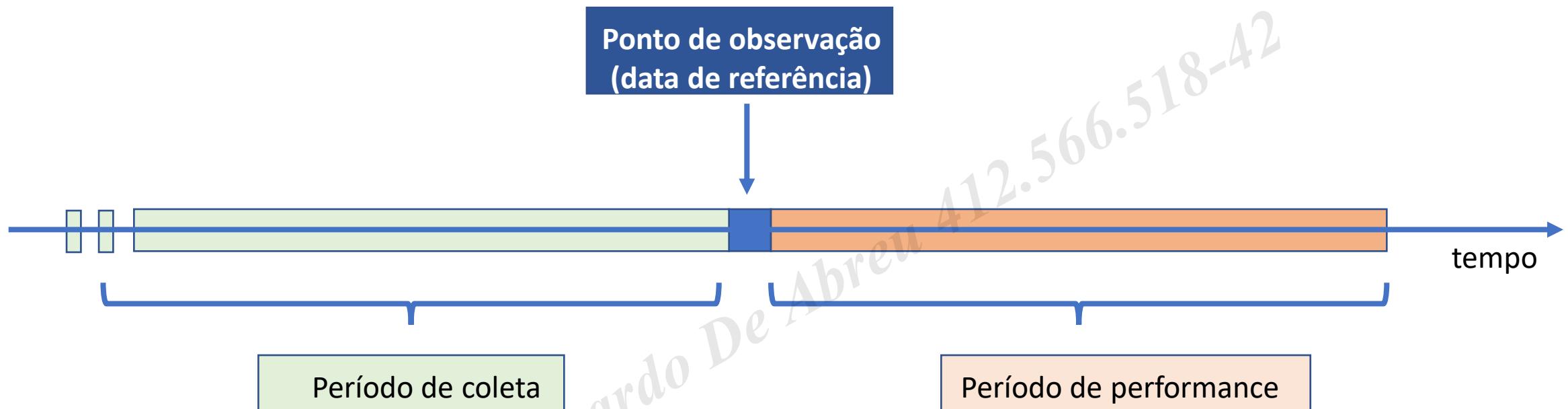
*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor. Proibida a reprodução, total ou parcial, sem autorização. Lei nº 9610/98



Modelos preditivos

Como é isso?

Desenho de safra (ou coorte)



Exemplo de desenho amostral para modelo preditivo

Estrutura da base de dados

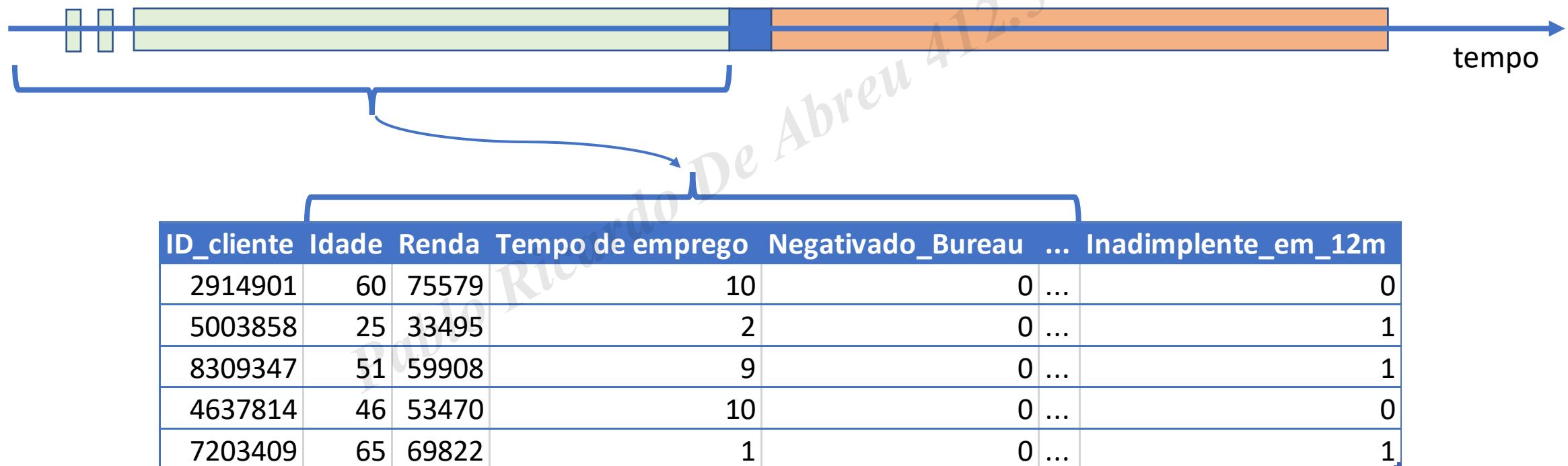
Cada linha
representa uma
observação

Cada coluna é uma variável

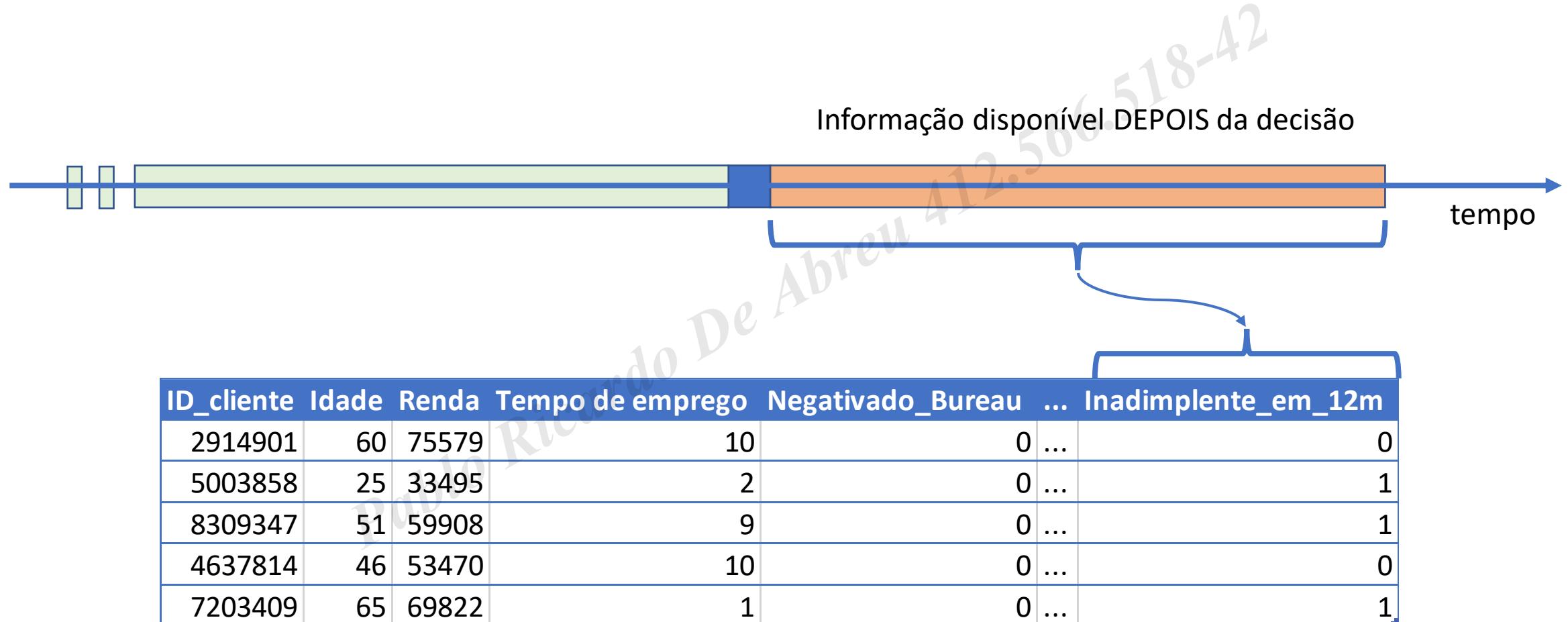
ID_cliente	Idade	Renda	Tempo de emprego	Negativado_Bureau	...	Inadimplente_em_12m
2914901	60	75579		10	0 ...	0
5003858	25	33495		2	0 ...	1
8309347	51	59908		9	0 ...	1
4637814	46	53470		10	0 ...	0
7203409	65	69822		1	0 ...	1

Desenho amostral

Informações disponíveis ANTES da decisão.



Desenho amostral



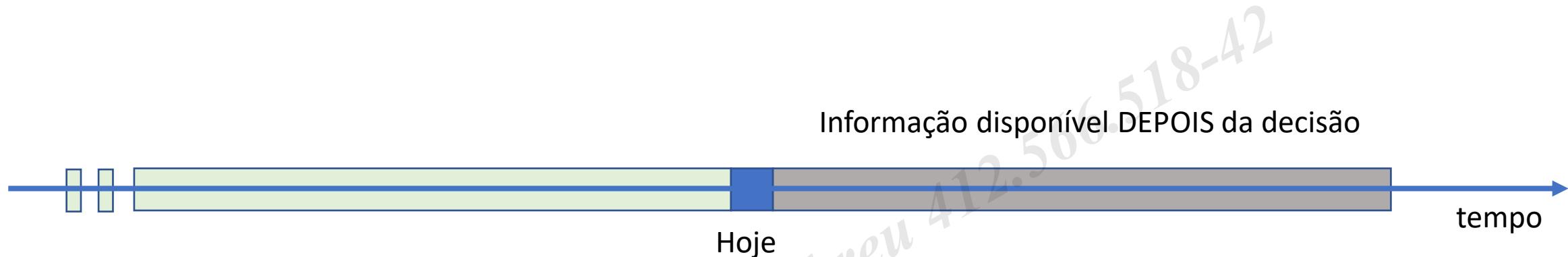
Previsão do modelo - *treinamento*

$$PD = F(X_1, X_2, \dots, X_n)$$

Previsão – probabilidade de inadimplência

ID_cliente	Idade	Renda	Tempo de emprego	Negativado_Bureau	...	Inadimplente_em_12m	Prob(Inad_12)
6046023	63	44095		7		1 ...	1 64,2%
6953168	45	47550		8		1 ...	1 63,3%
8529261	58	89313		7		1 ...	1 71,3%
1281888	42	58532		7		0 ...	0 15,3%
4540759	37	80840		2		0 ...	1 86,7%

Previsão do modelo - *aplicação*



ID_cliente	Idade	Renda	Tempo de emprego	Negativado_Bureau	...	Inadimplente_em_12m	Prob(Inad_12)
2064492	57	78774		10	0 ...		26,4%
8858083	63	99559		4	1 ...		39,4%
6937124	41	49791		2	0 ...		93,6%
7352807	42	15314		8	0 ...		34,1%
5088775	31	73199		4	0 ...		97,4%

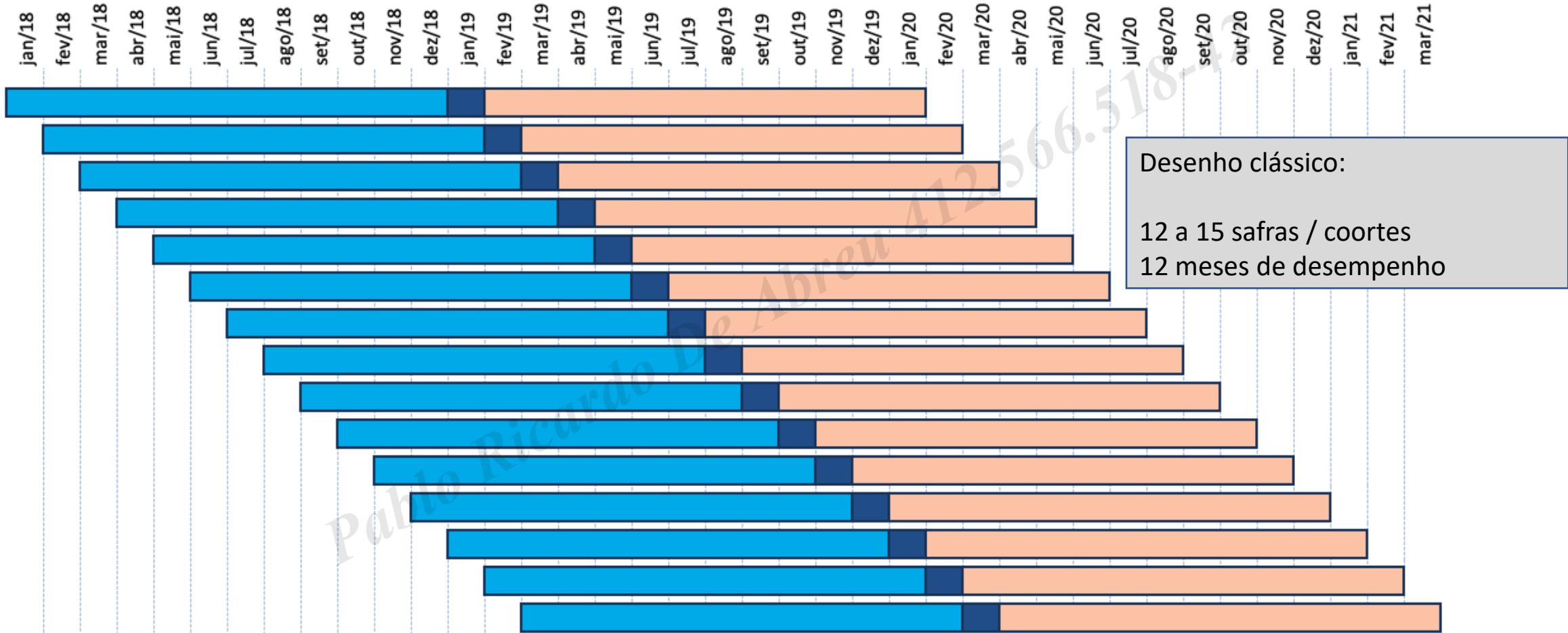
Na aplicação, não temos a marcação de inadimplência, mas temos a probabilidade dada pelo modelo.
Essa é a PREVISÃO do modelo.



Desenho clássico

- E se o mês de observação sofre algum efeito sazonal como final de ano etc?
- Para isto, se faz de observar 12 meses de referência diferentes (chamados de “safras” ou “coortes”)

Desenho do modelo



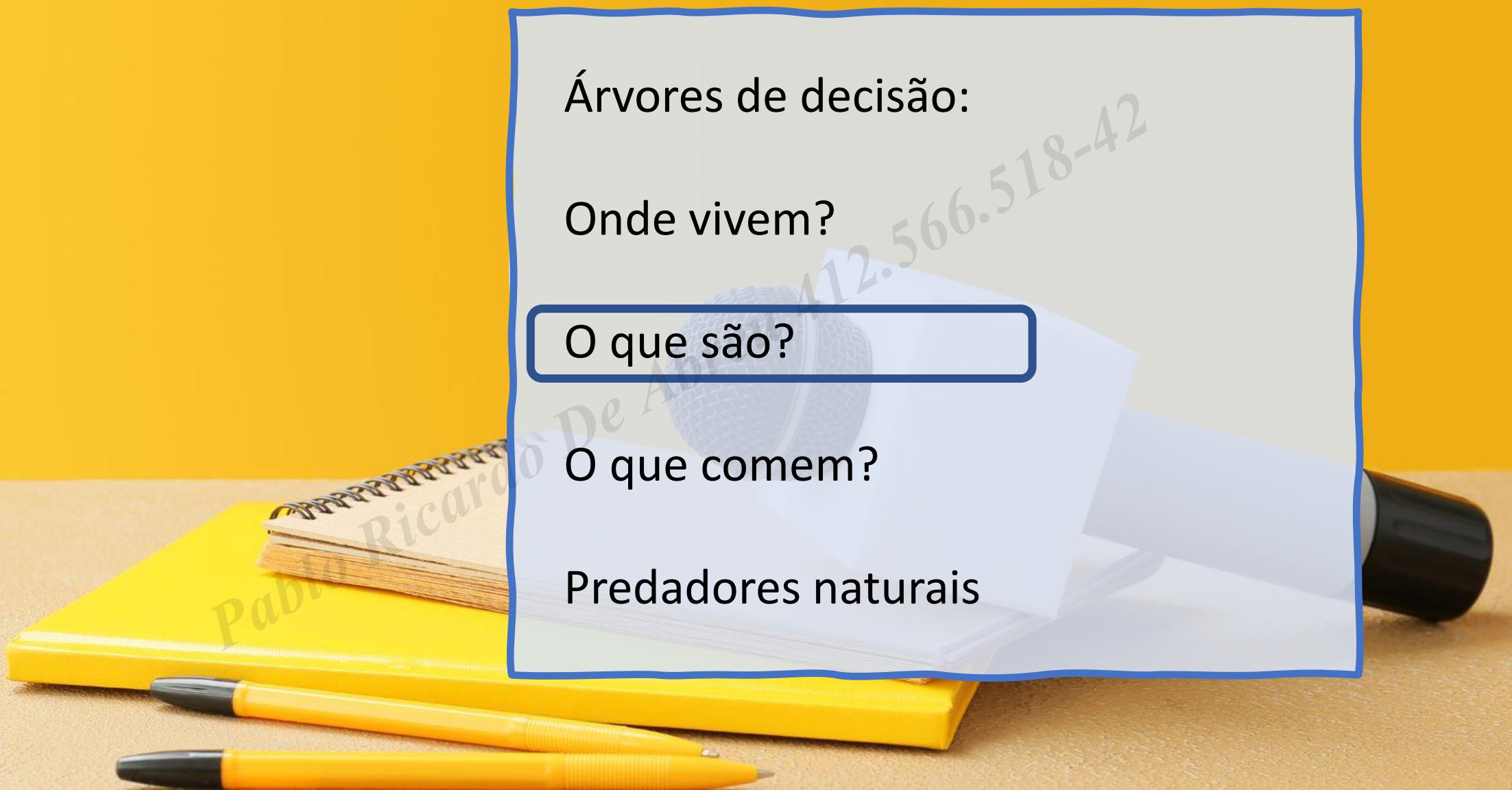
Árvores de decisão:

Onde vivem?

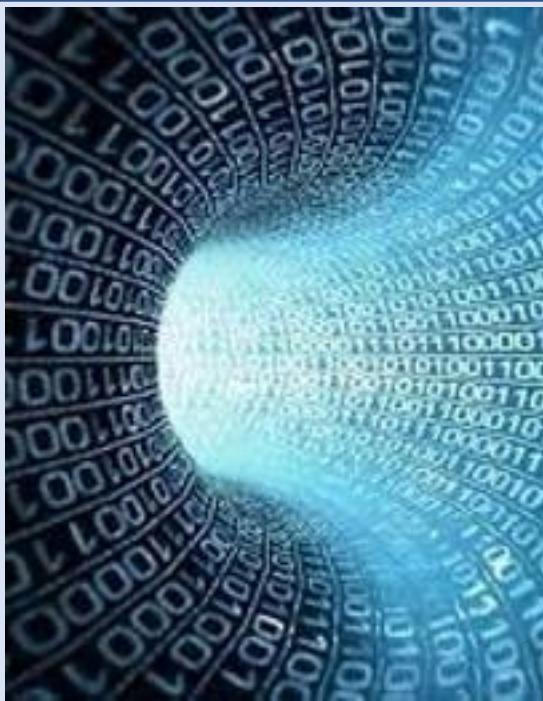
O que são?

O que comem?

Predadores naturais



Classificação dos algoritmos



Paradigma Machinelârnico

- Árvores de decisão
- Bagging
- Boosting
- K-NN
- Redes Neurais
- Support Vector Machines



Paradigma Estatístico

- Regressão
- GLM
- GLMM
- ANOVA

Estamos aqui!

Machine learning vs estatística

- Machine learning

- Apenas busca padrões nos dados
- Busca que sejam generalizáveis
- Mede o erro de forma pragmática



- Paradigma estatístico

- Supõe estrutura probabilística
- Avalia as suposições
- **Faz inferência**



Classificação dos algoritmos



Supervisionados

- Regressão
- GLM
- GLMM
- Support vector machines
- Naive Bayes
- K-nearest neighbors
- Redes Neurais
- Decision Trees



Não supervisionados

- K-Means
- Métodos hierárquicos
- Mistura Gaussiana
- DBScan
- Mini-Batch-K-Means

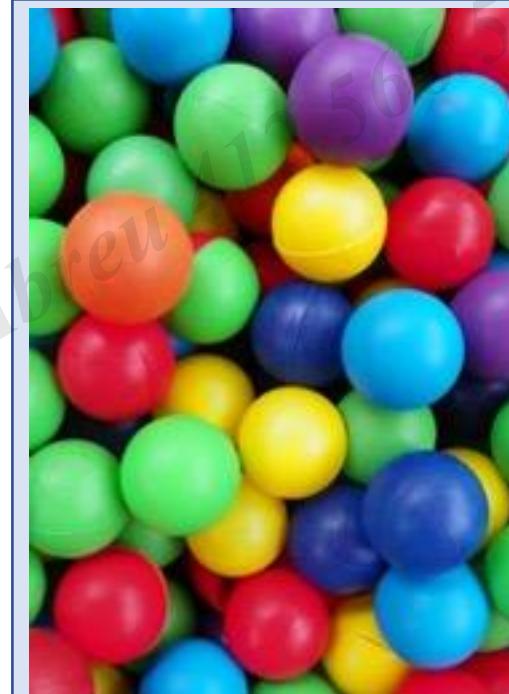
Estamos aqui!

Classificação dos algoritmos



Resposta contínua

- Regressão
- GLM
- GLMM
- Support vector machines
- K-nearest neighbors
- Redes Neurais
- Regression Trees



Resposta discreta

- Regressão logística
- Classification trees
- Redes Neurais
- GLM
- GLMM

Estamos aqui!



Nosso problema: classificar sobreviventes

Imagen: https://commons.wikimedia.org/wiki/File:Sea_Trials_of_RMS_Titanic,_2nd_of_April_1912.jpg

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.
Proibida a reprodução, total ou parcial, sem autorização. Lei nº 9610/98



Reflexões sobre a base de dados

População

- ~ 2.200 pessoas
- ~ 1.300 passageiros
- Mais de 1.500 mortos

Amostra

- 891 pessoas
- 549 não sobreviventes
- 342 sobreviventes



Objetivos do algoritmo de Machine Learning

- Classificar da melhor forma possível a variável resposta
 - ... Através de segmentações
 - ... Usando as variáveis explicativas
- Obter insights
 - ... Das relações entre a variável resposta e as explicativas
 - ... Explorar interações

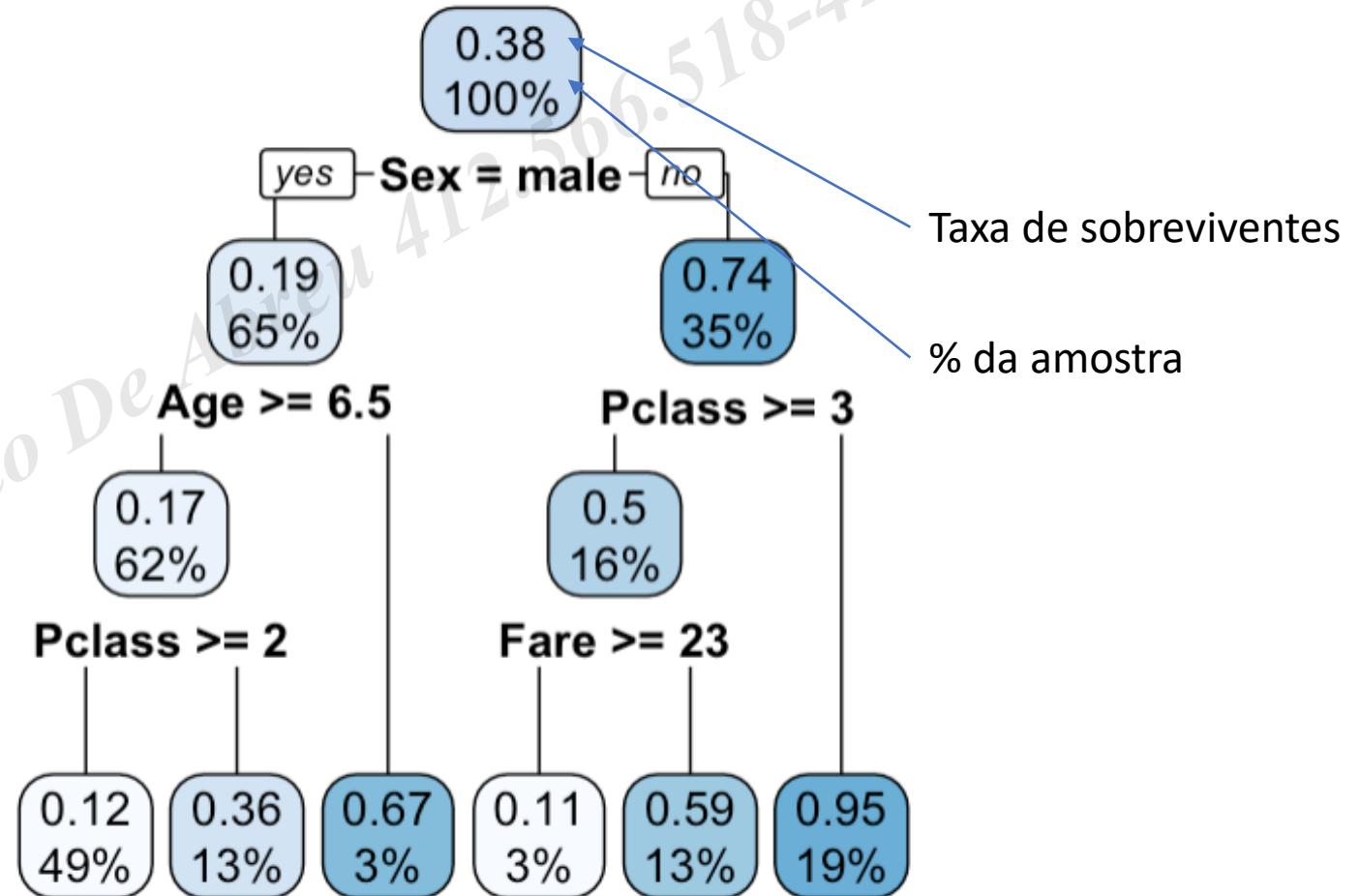


OMML1_script01-Primeiro_contato_com_arvores.R

O que é uma árvore de decisão?

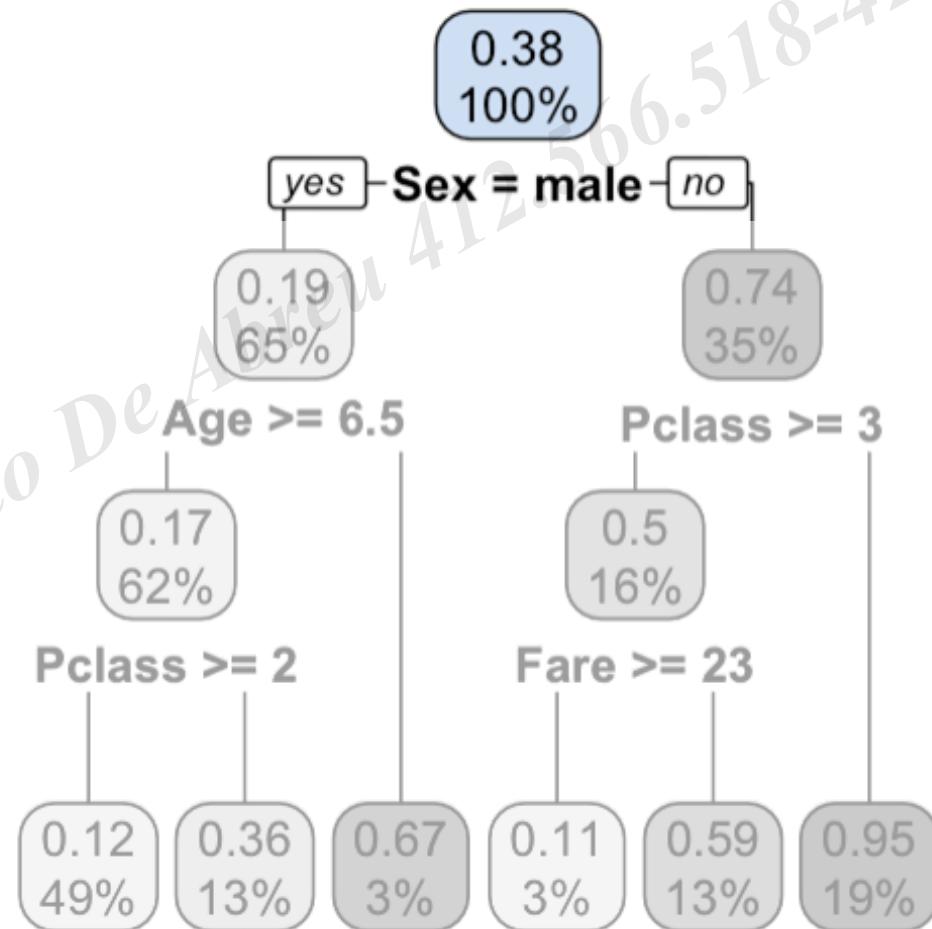
A árvore de decisão é:

Uma sequência de segmentações binárias
Que visa homogeneidade da variável resposta



O que é uma árvore de decisão?

Inicialmente temos 891 passageiros dos quais
342 sobreviveram (38%)
549 não sobreviveram

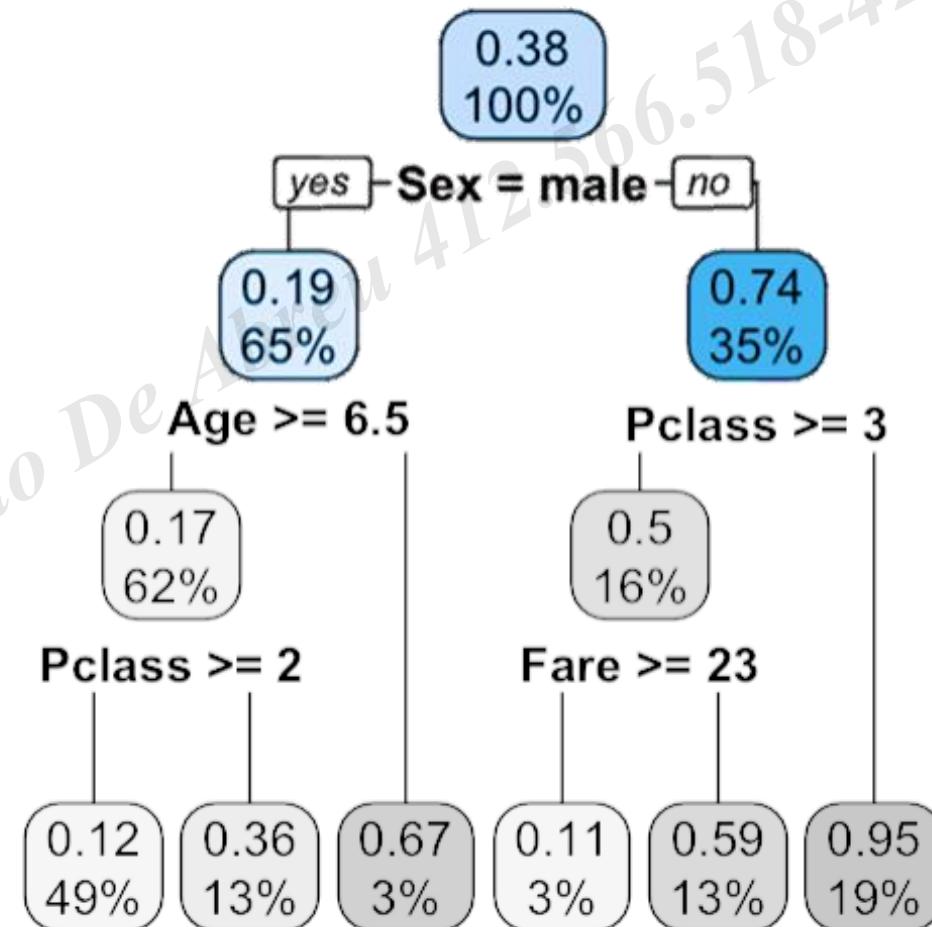


O que é uma árvore de decisão?

Dos 891, podemos segmentá-los em:

577 homens (65%) dos quais
109 sobreviveram (19%)
468 não sobreviveram

314 mulheres (35%) das quais
233 sobreviveram (74%)
81 não sobreviveram



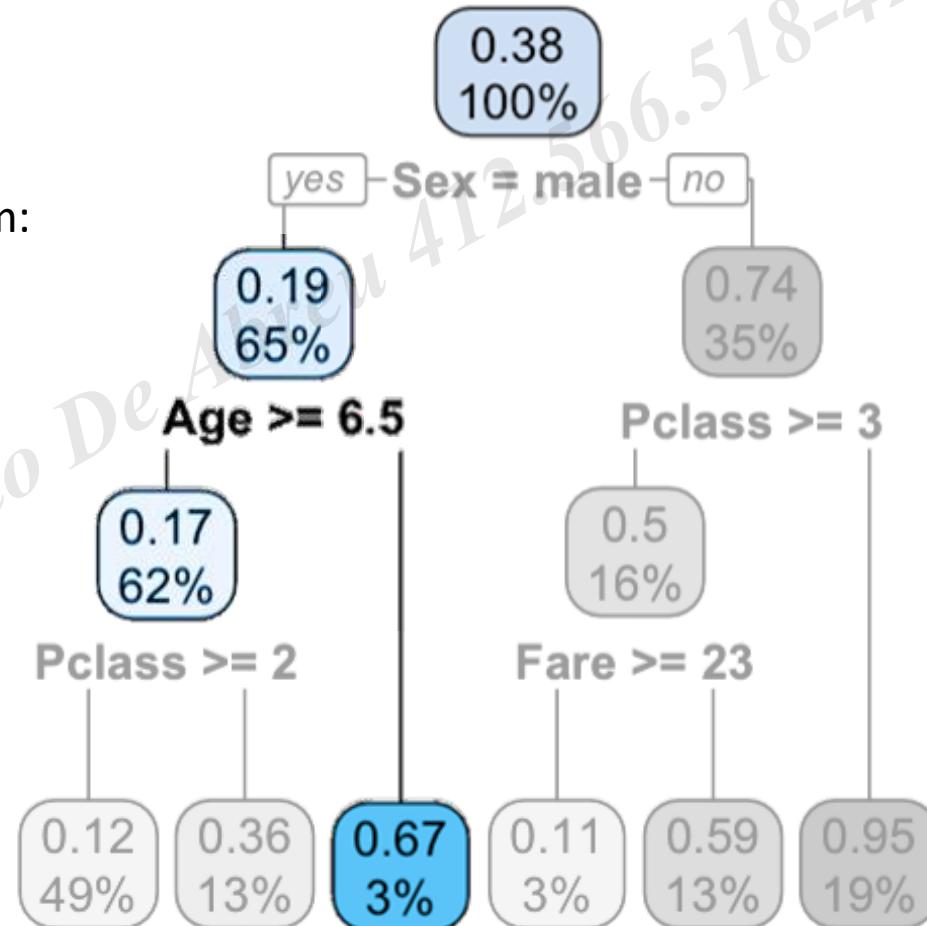
O que é uma árvore de decisão?

Dos 891, podemos segmentá-los em:

577 homens que por sua vez segmentamos em:

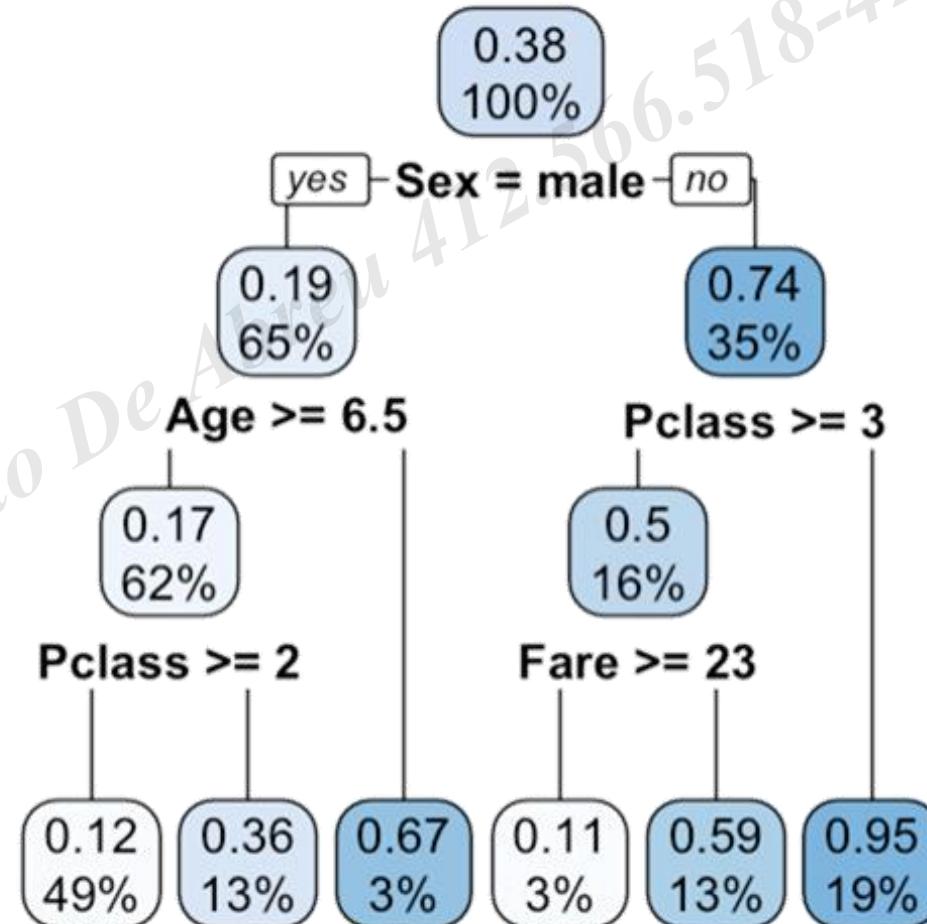
24 crianças (< 6,5 anos) das quais
16 sobreviveram (67%)
8 não sobreviveram

533 adultos ($\geq 6,5$ anos) dos quais
93 sobreviveram (17%)
553 não sobreviveram



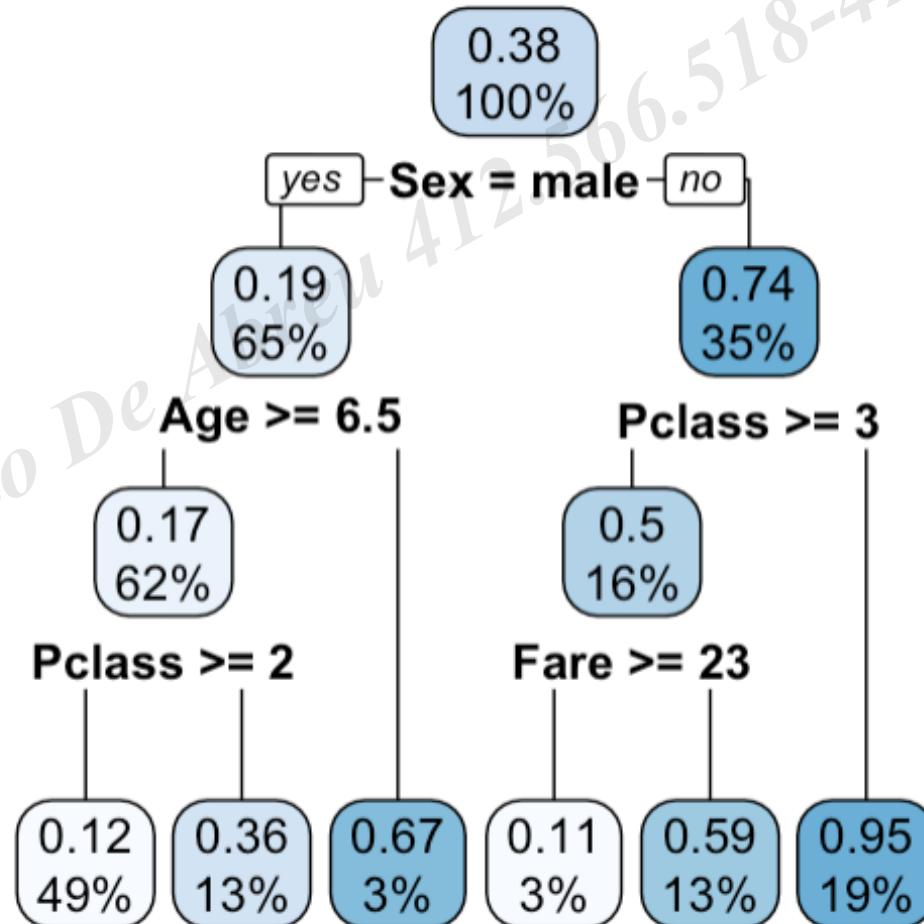
O que é uma árvore de decisão?

E assim continuamos a “requebrar” a amostra até “não valer a pena” fazer mais quebras.



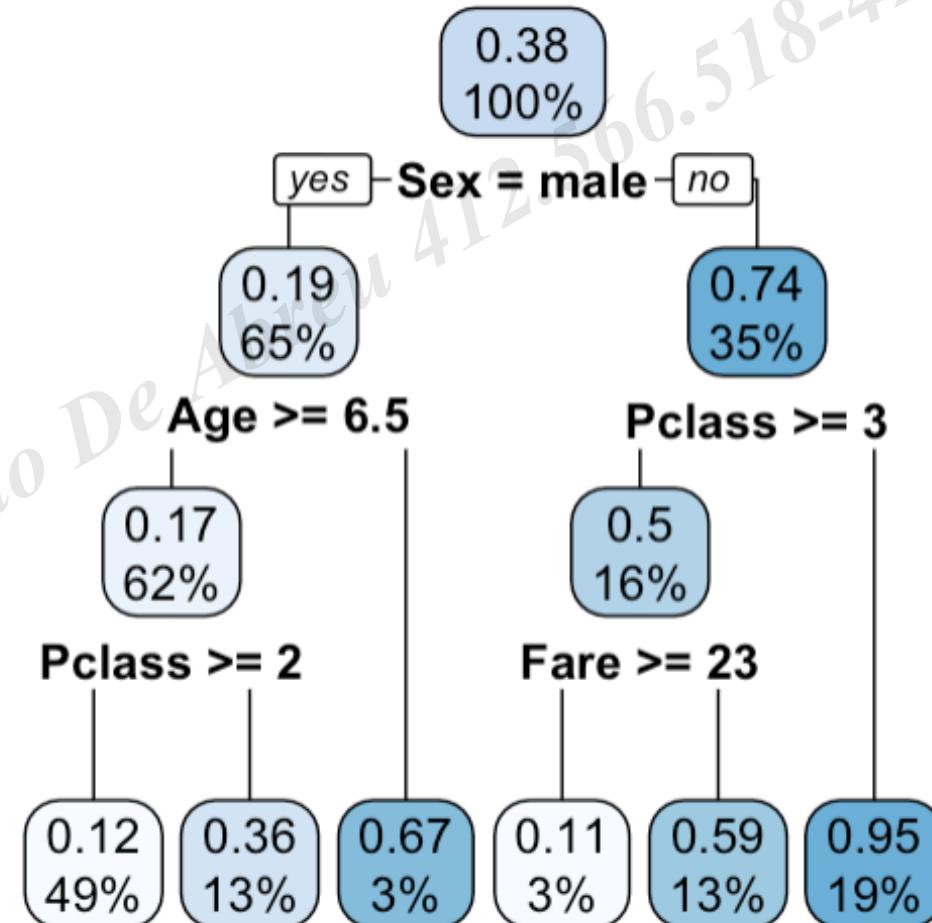
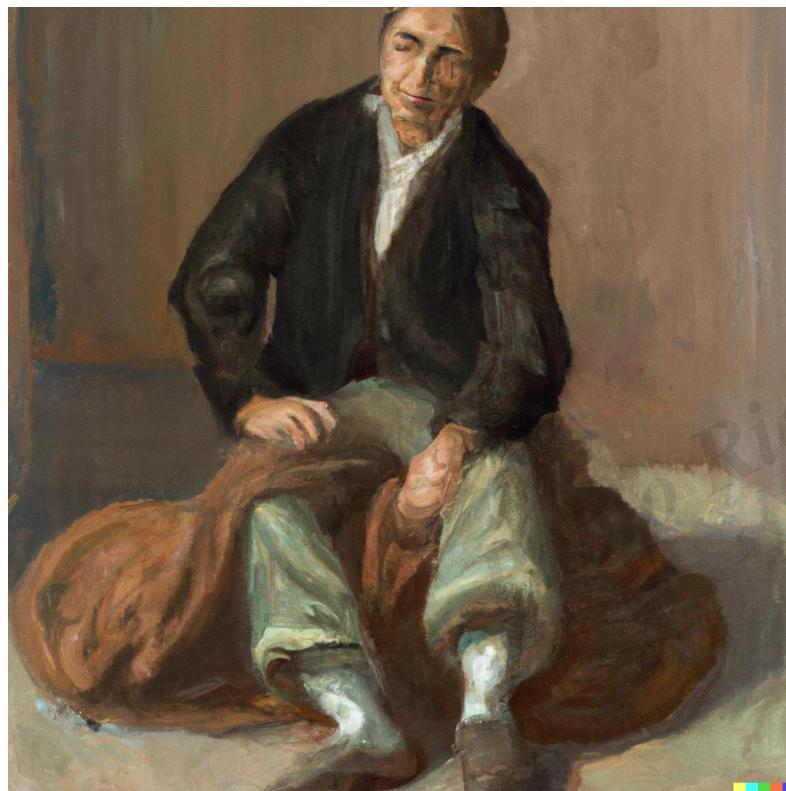
Vamos experimentar??

Vamos classificar este garoto:



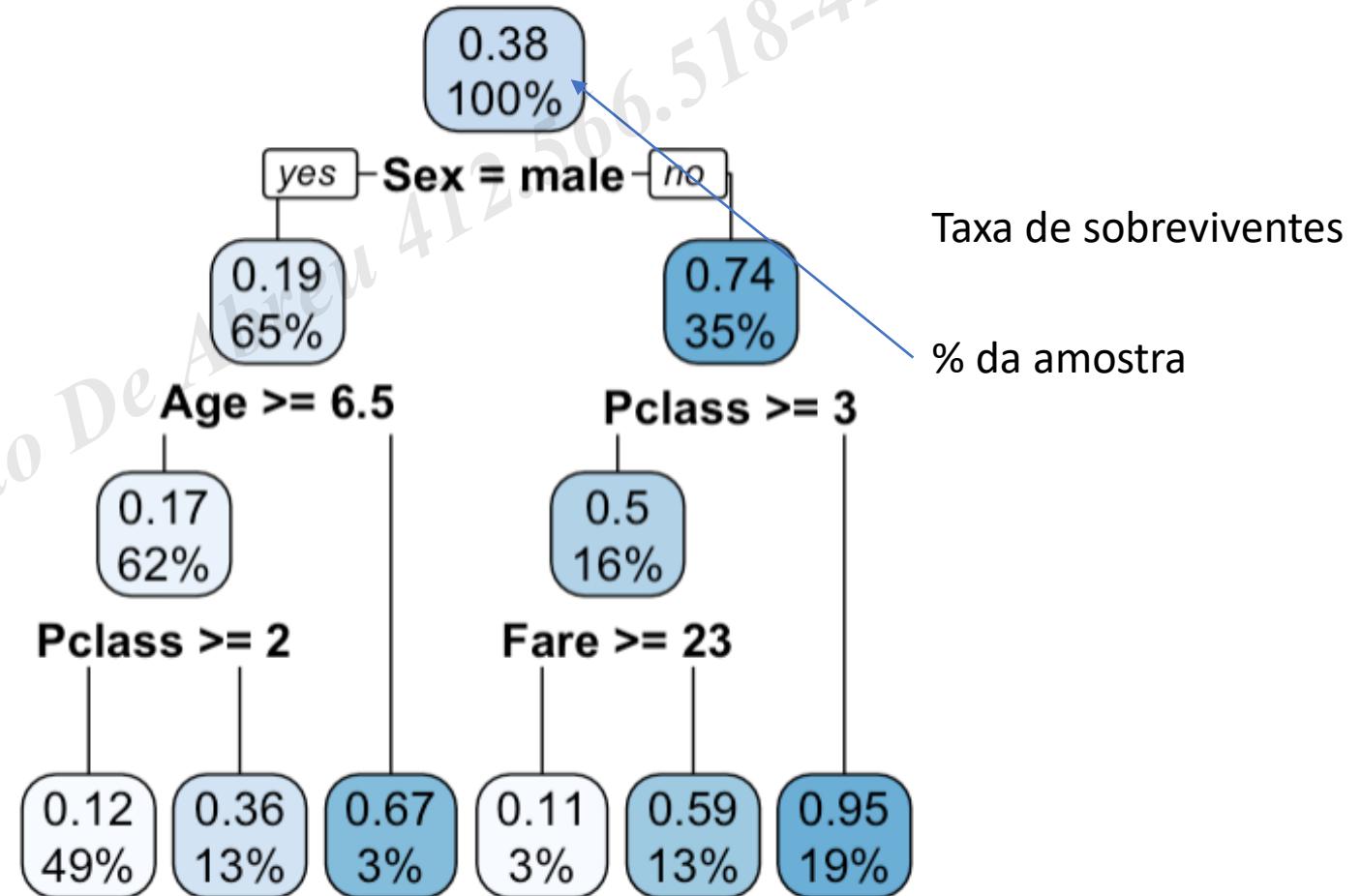
Vamos experimentar??

Vamos classificar este pobre homem:



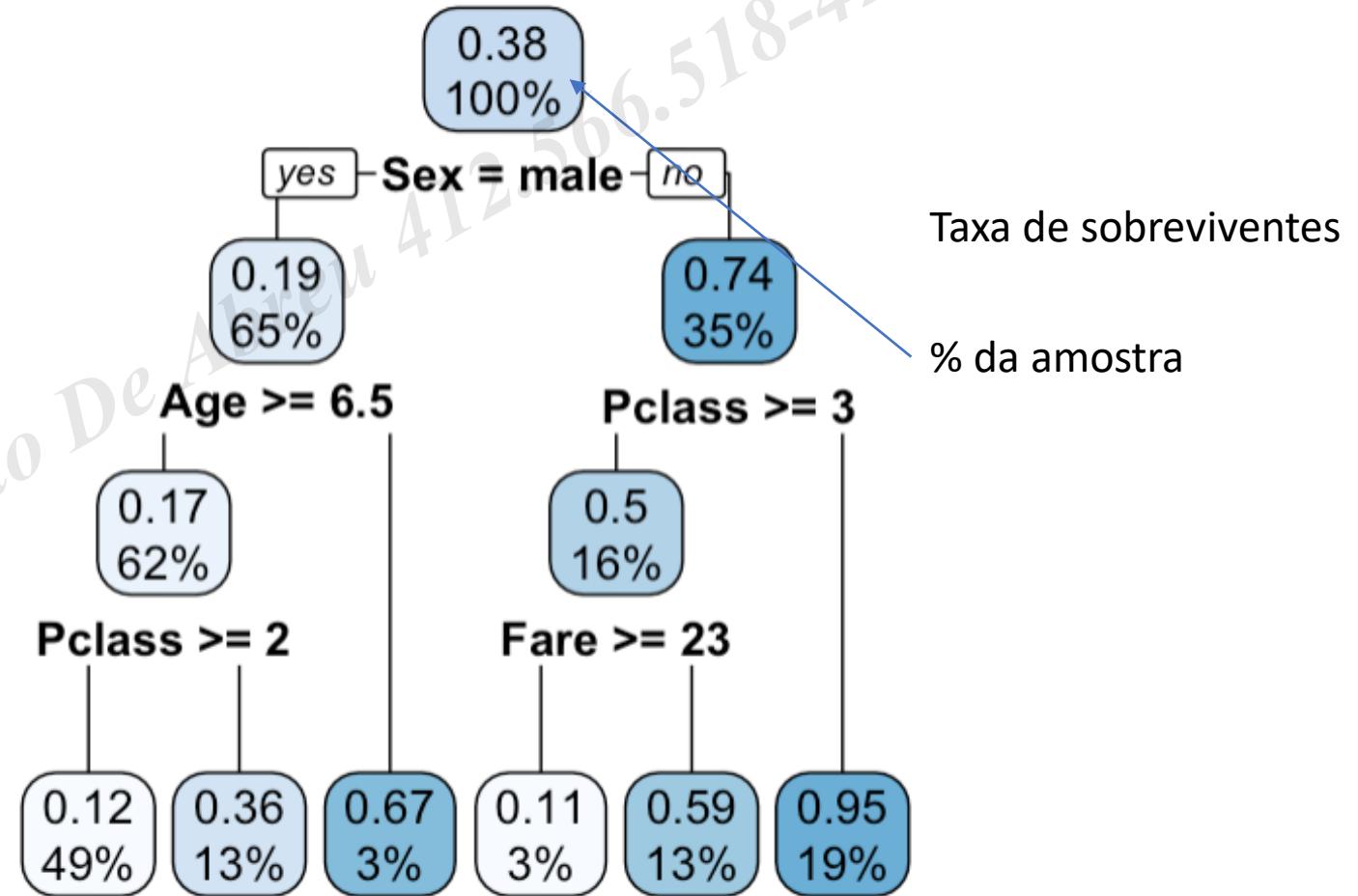
Vamos experimentar??

Vamos classificar este rico senhor:



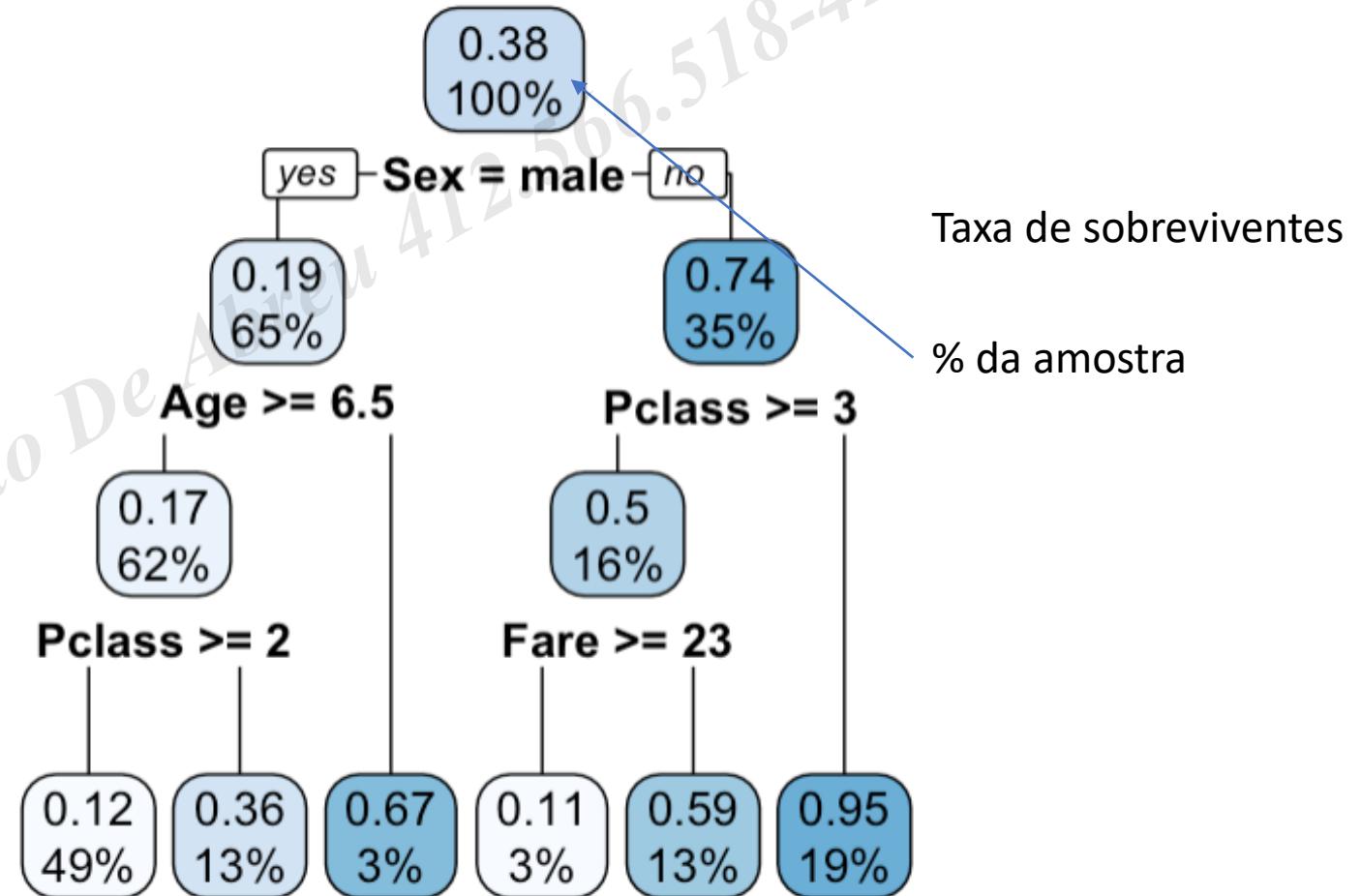
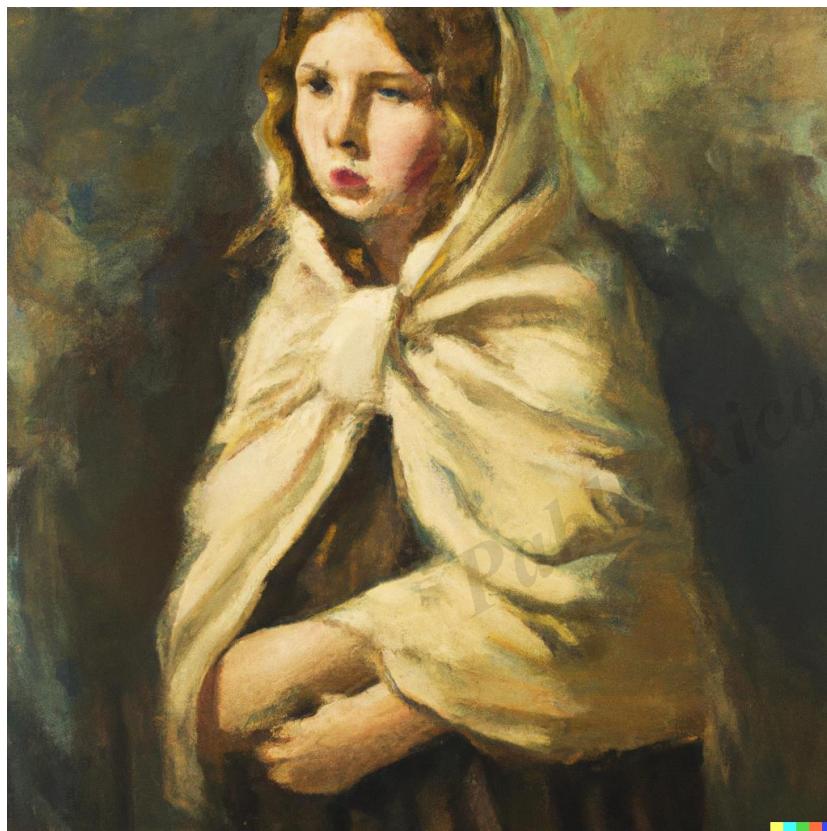
Vamos experimentar??

Vamos classificar este rica senhora:



Vamos experimentar??

Vamos classificar esta pobre moça:



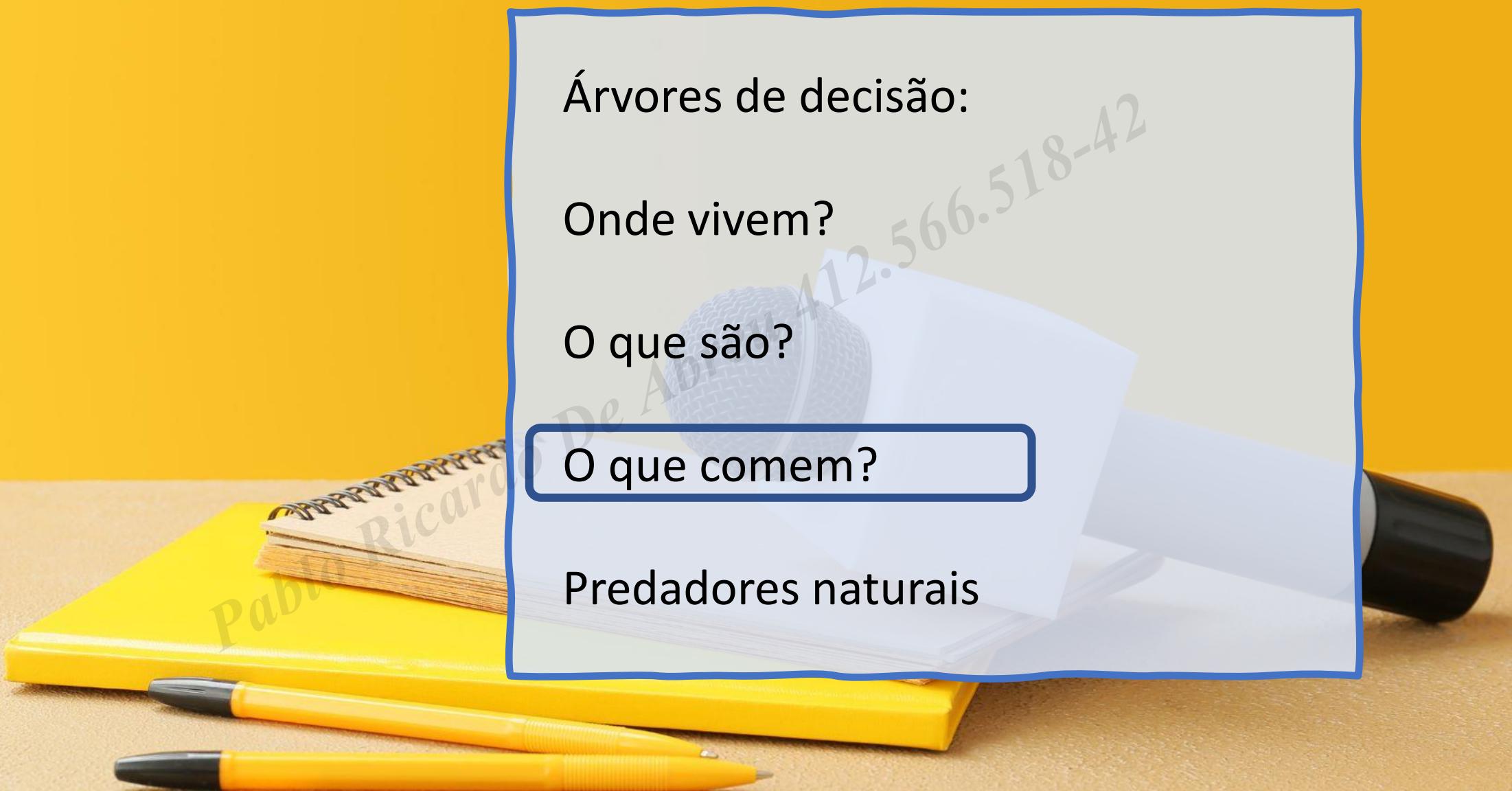
Árvores de decisão:

Onde vivem?

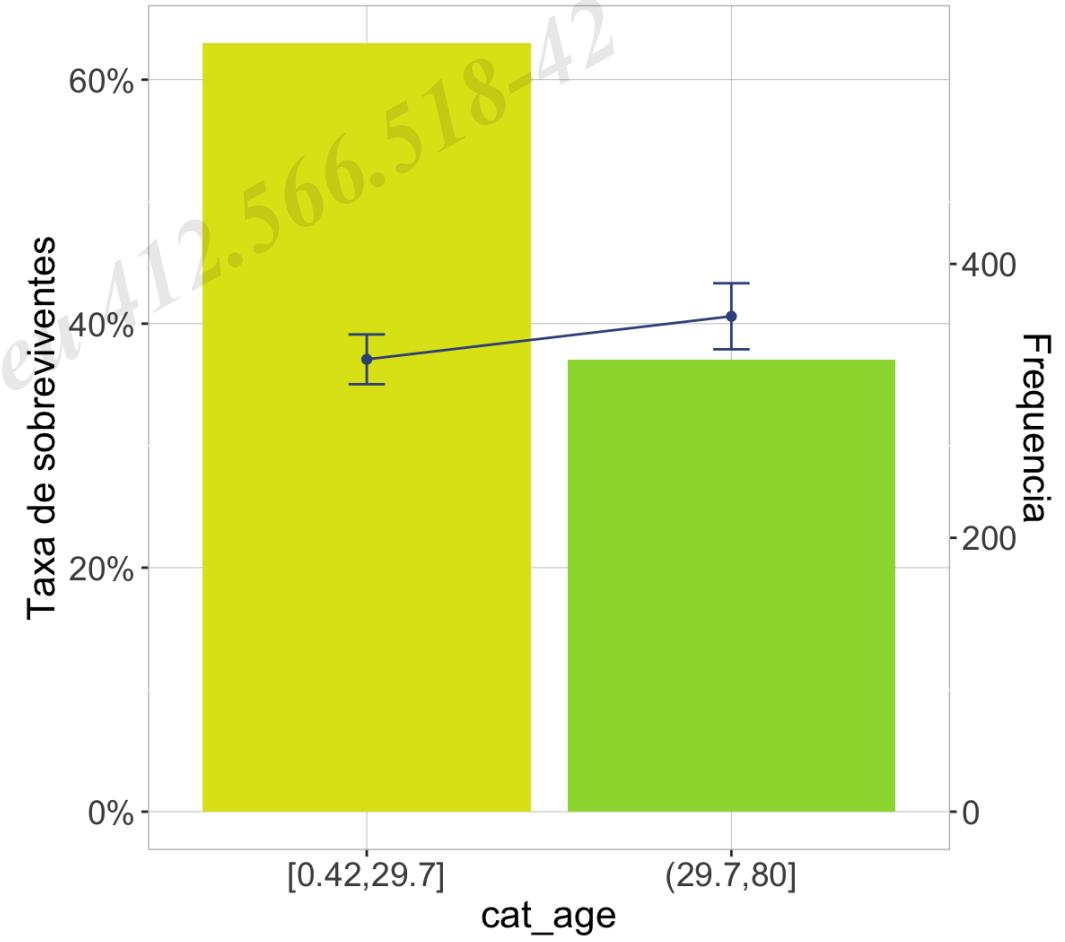
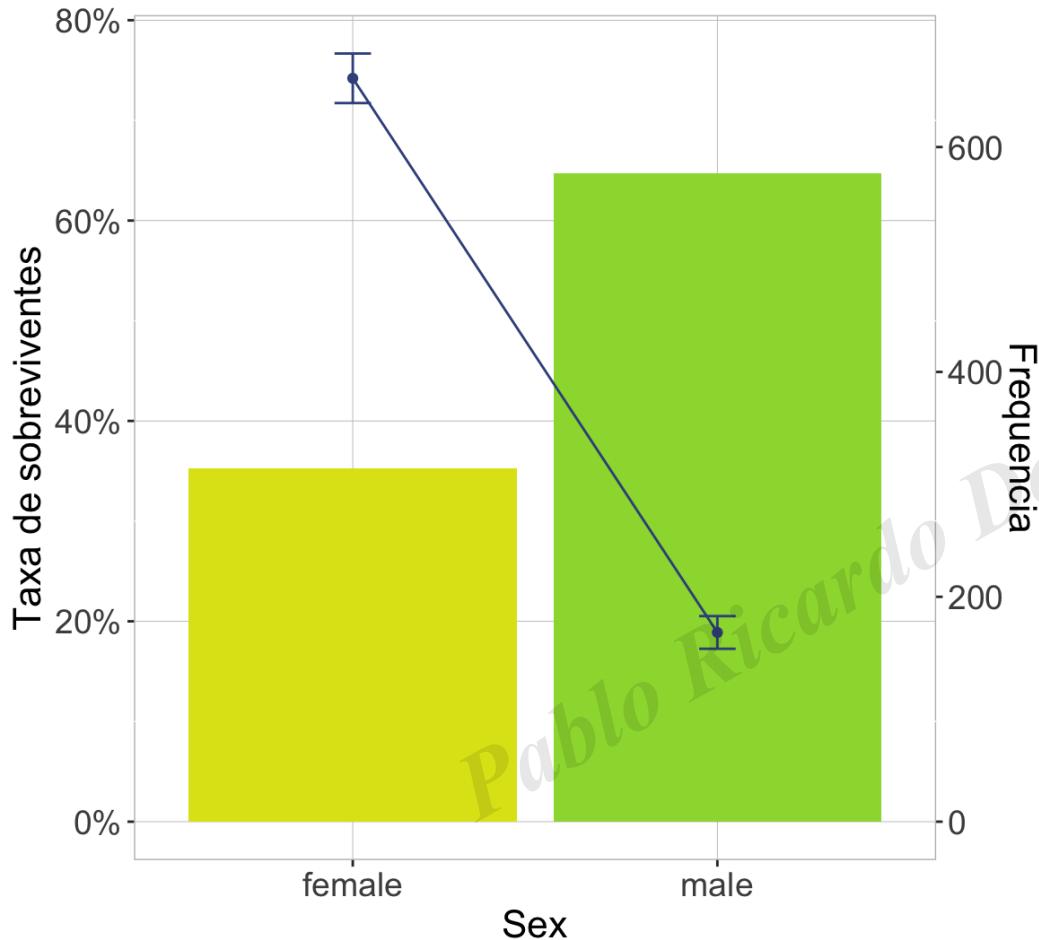
O que são?

O que comem?

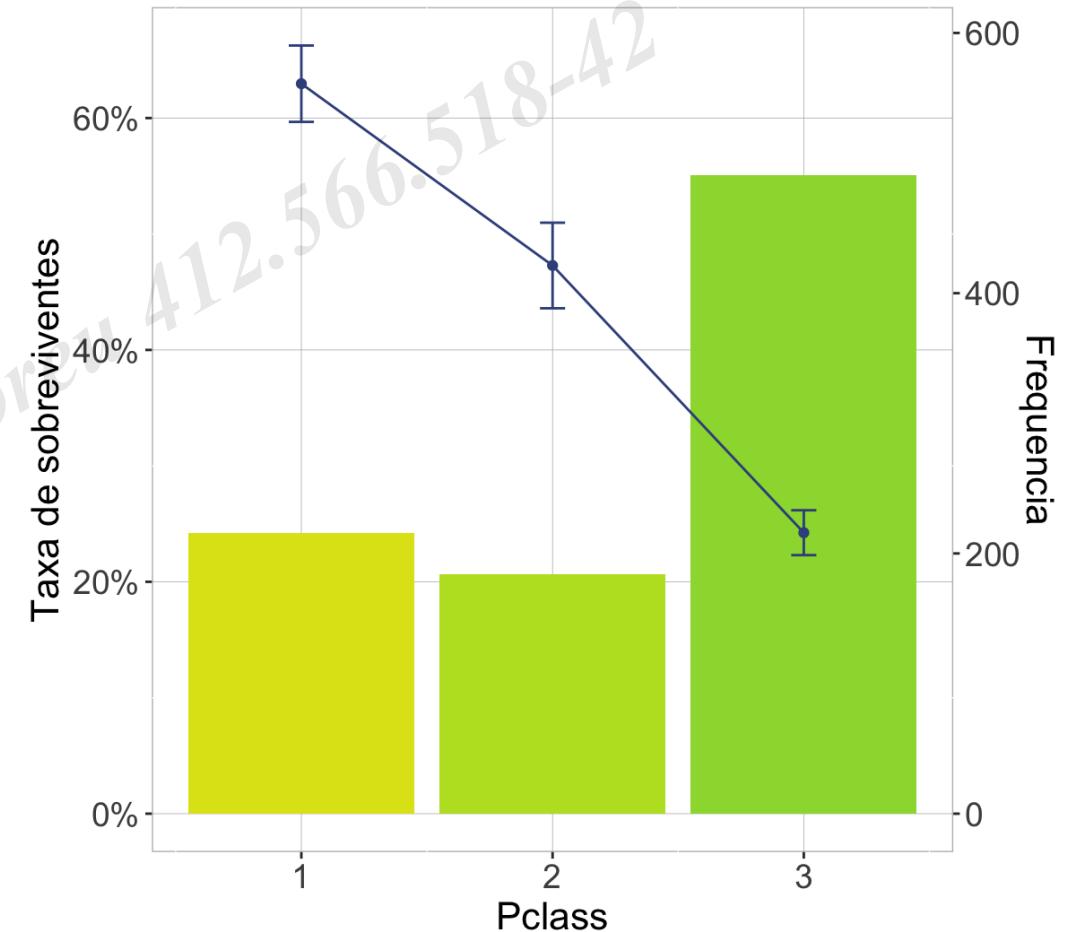
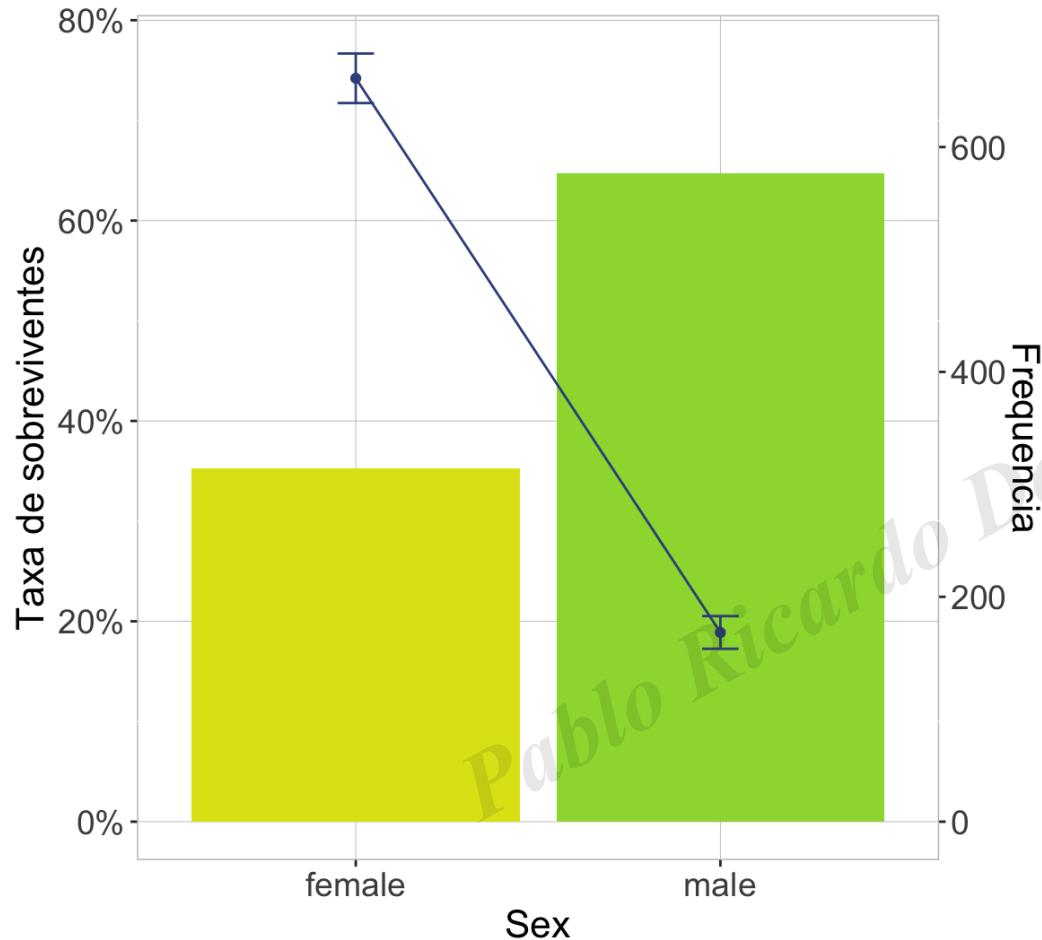
Predadores naturais



Qual é ‘melhor’?



E agora?



Medida de “informação”

- Precisamos de uma métrica para ajudar a decidir qual variável agrupa mais informação.
- Vamos dizer que se todas as observações são idênticas em um grupo, esse grupo tem “pureza” total.
- Quanto mais heterogêneo o grupo, maior será a “impureza”.
- Vamos então definir quantitativamente o que é “impureza”:

Definições de impureza

- Gini
- Entropia de Shannon

Como a árvore encontra a melhor quebra?
Com uma métrica de 'impureza'

Índice de Gini

$$I_g(p) = 1 - \sum_{i=1}^J p_i^2$$

- Impureza máxima com distribuição uniforme
- Impureza mínima na concentração total

Entropia

$$H = - \sum_{i=1}^J p_i \log_2(p_i)$$

Ganho de informação:

$$GI(T, a) = H(T) - H(T|a)$$

- Impureza máxima com distribuição uniforme
- Impureza mínima na concentração total

Algoritmo básico

1. Para cada variável, buscar a melhor regra binária
2. Escolher aplicar melhor segmentação dentre todas as variáveis
3. Recursivamente, para cada folha, repetir os passos 1 e 2 até que uma regra de parada seja atingida

Implementação web interativa:

<https://rawgit.com/longhowlam/titanicTree/master/tree.html>

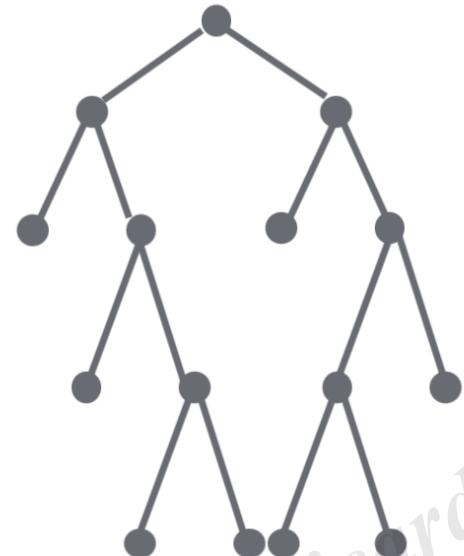


Hiperparâmetros

São parâmetros que controlam o algoritmo como:

1. Número mínimo de observações por folha
2. Profundidade máxima
3. CP – Custo de complexidade

Custo de complexidade

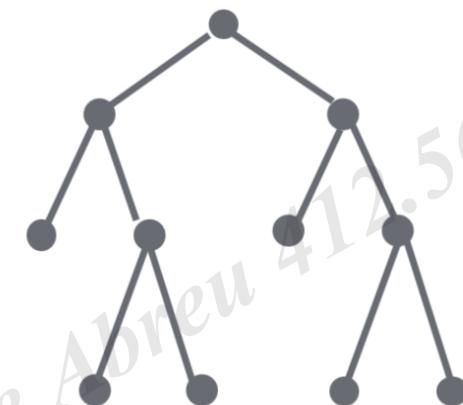


Custo de
Complexidade

Baixo

Complexidade
resultante

Alta



Médio

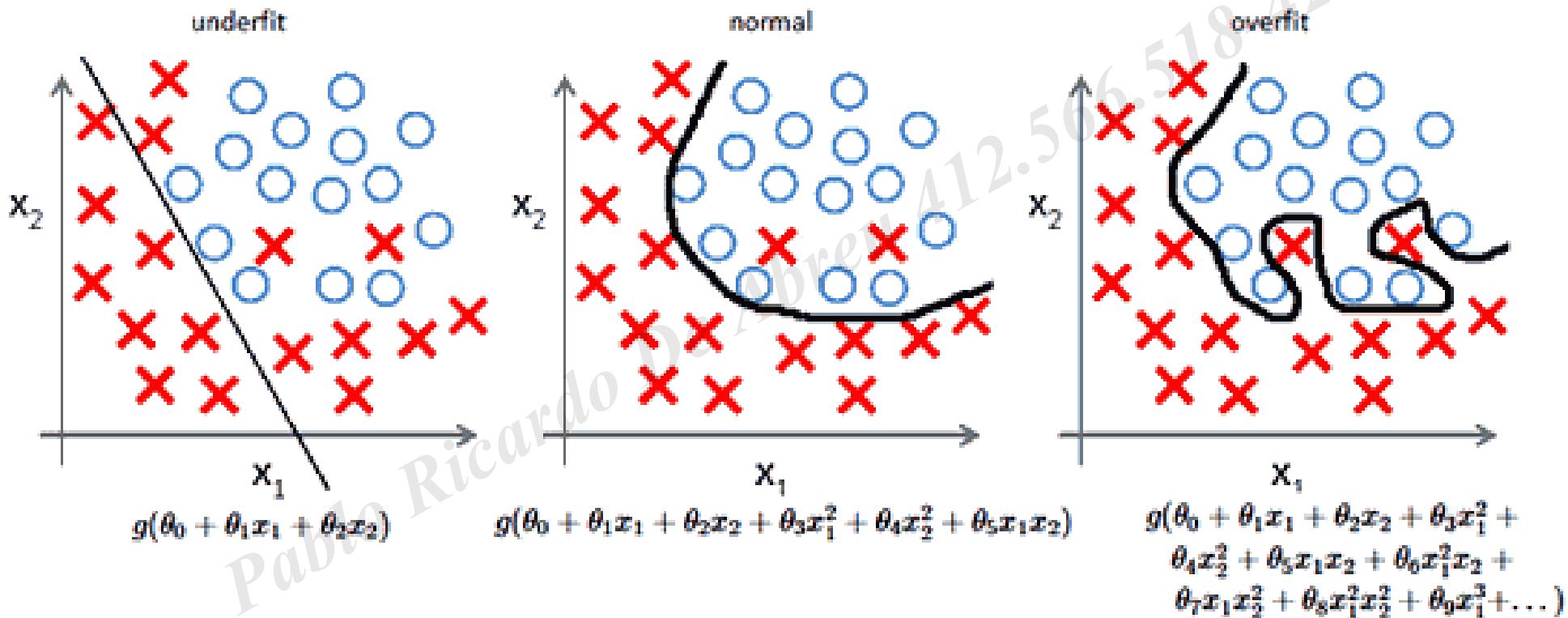
Alto



Média

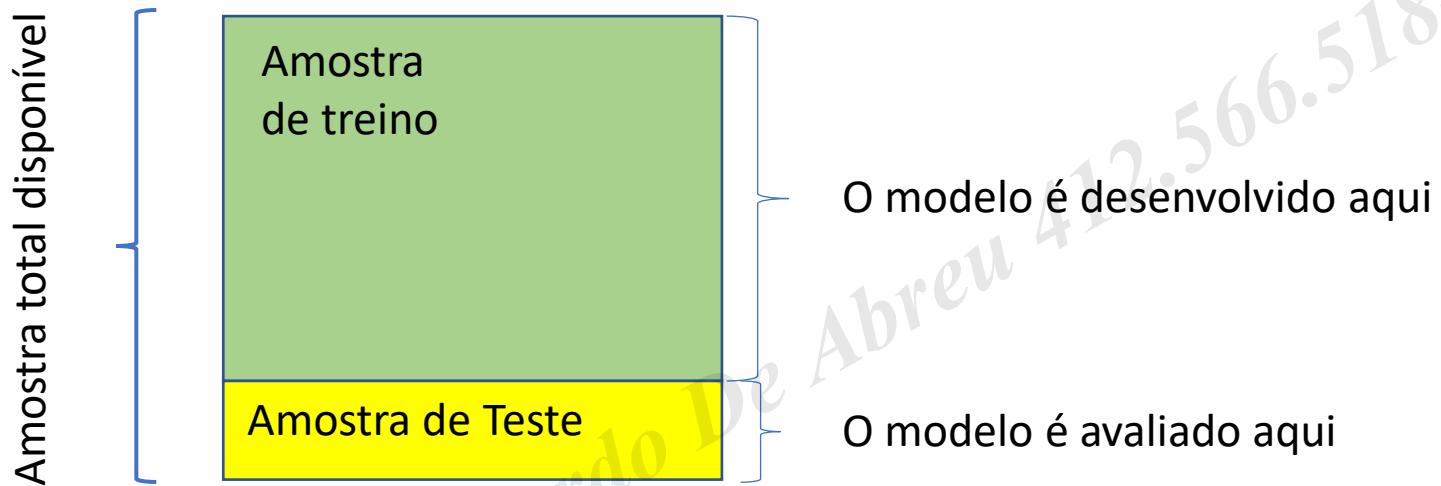
Baixa

Over fitting / Under fitting



<http://mlwiki.org/index.php/Overfitting>

Cross validation



A estratégia mais simples é dividir a base em treino e teste.
Desenvolvemos o modelo na base de treino e avaliamos na base de teste.

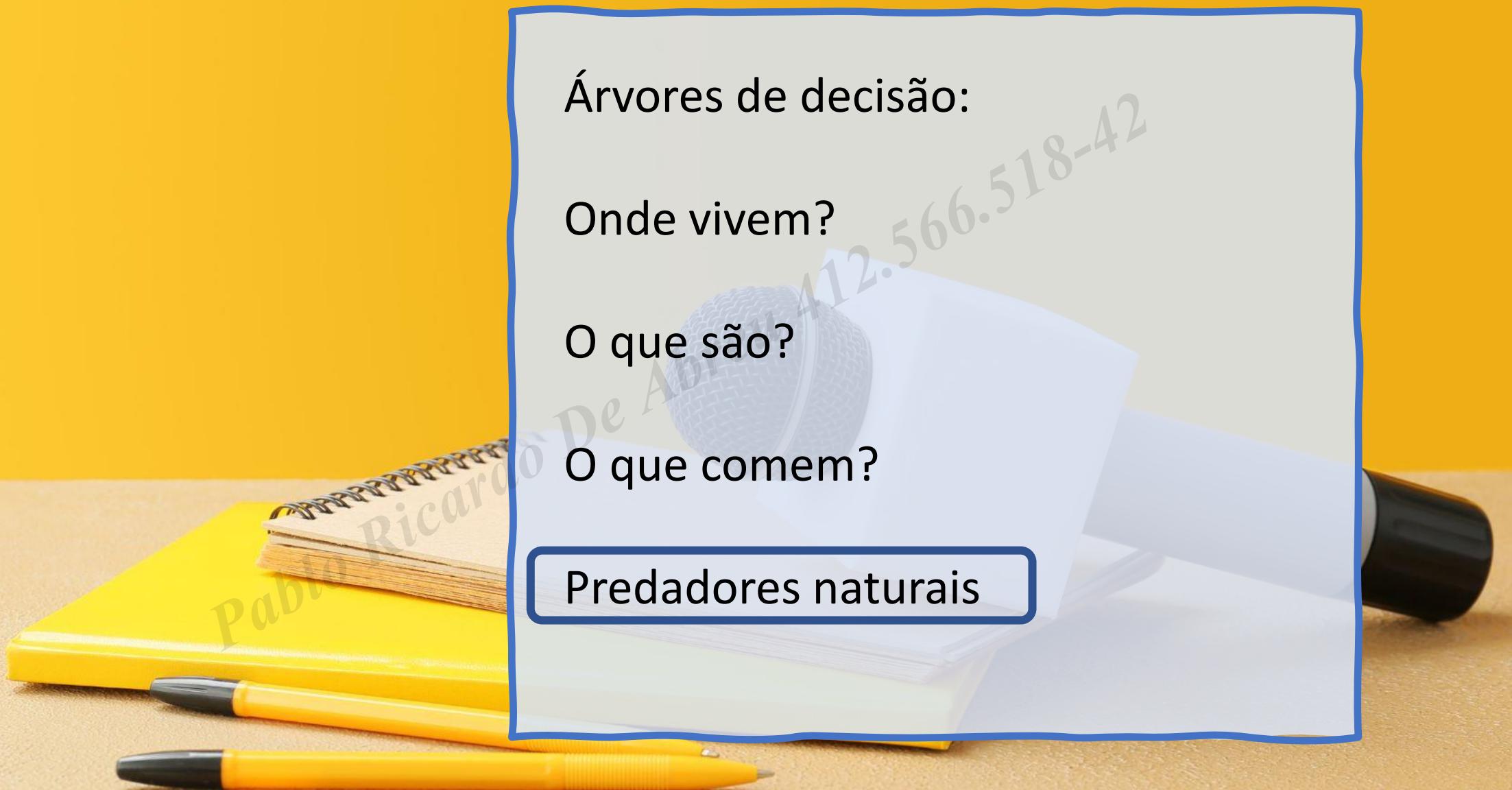
Árvores de decisão:

Onde vivem?

O que são?

O que comem?

Predadores naturais





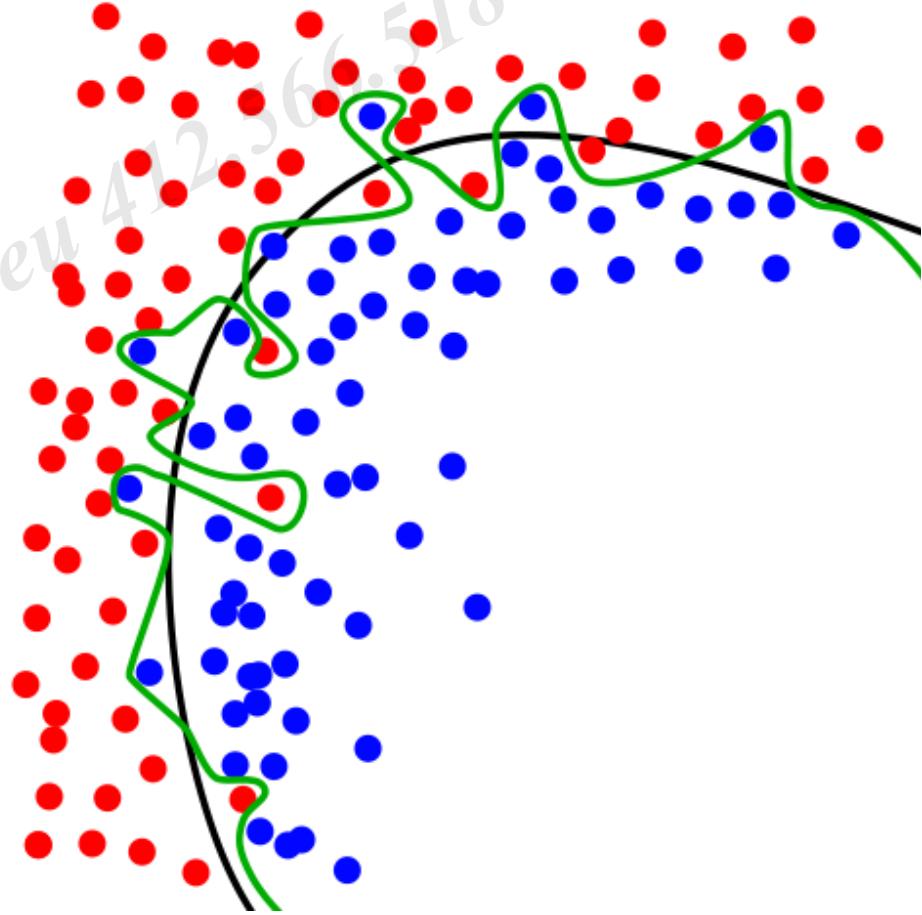
THE BEST WAY TO EXPLAIN OVERFITTING

O que é

Como evitar

Overfitting

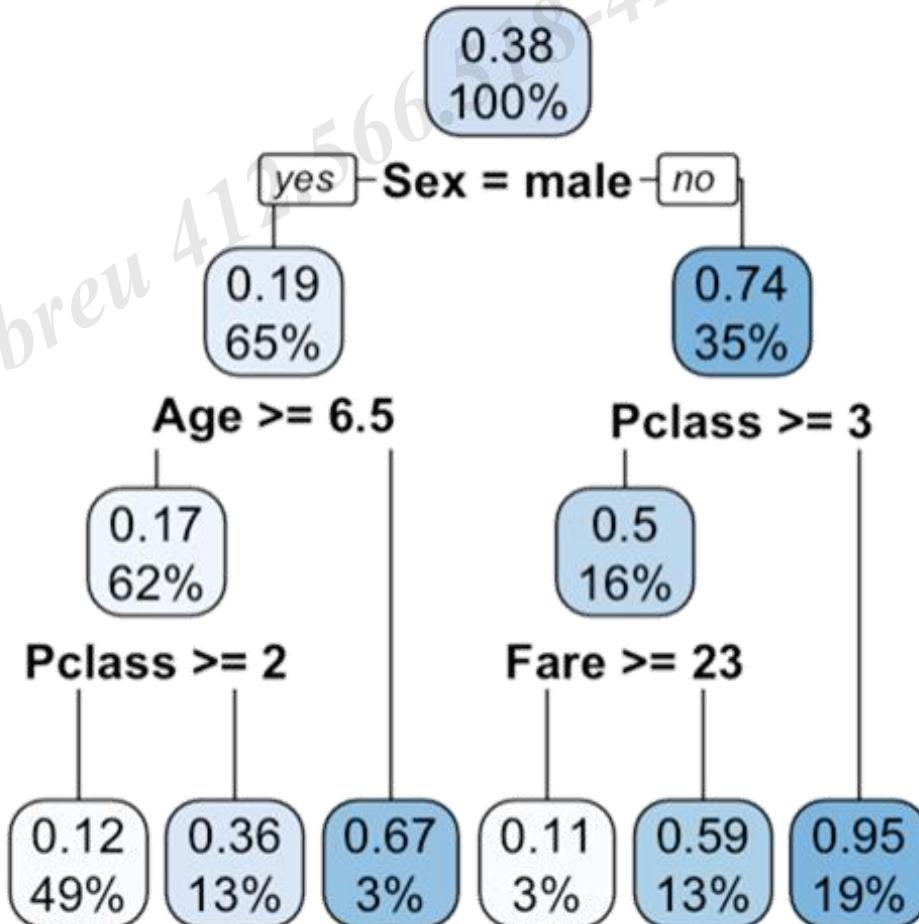
- *Overfitting* é o fenômeno do modelo se ajustando a características muito particulares de amostra que não se repetirão em outras amostras.
- Isso implica que a qualidade do modelo poderá ser bem menor na aplicação do modelo.



A árvore como um classificador

Requisitos:

Ter todas as variáveis.



A árvore como um classificador

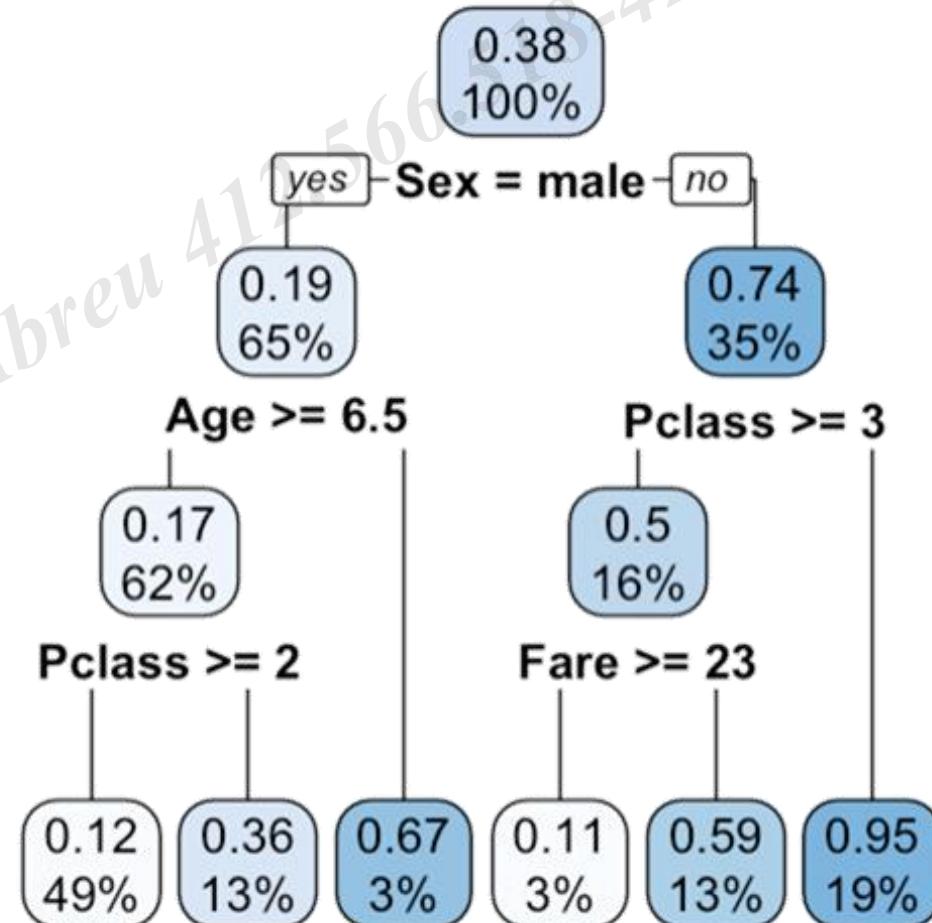
Probabilidade de evento da folha F:

$$P(S|F) = \frac{N_f^S}{N_f}$$

$P(S|F)$ - probabilidade de sucesso da folha F

N_f - é o número de indivíduos na folha F

N_f^S - é o número de sobreviventes na folha F



A árvore como um classificador

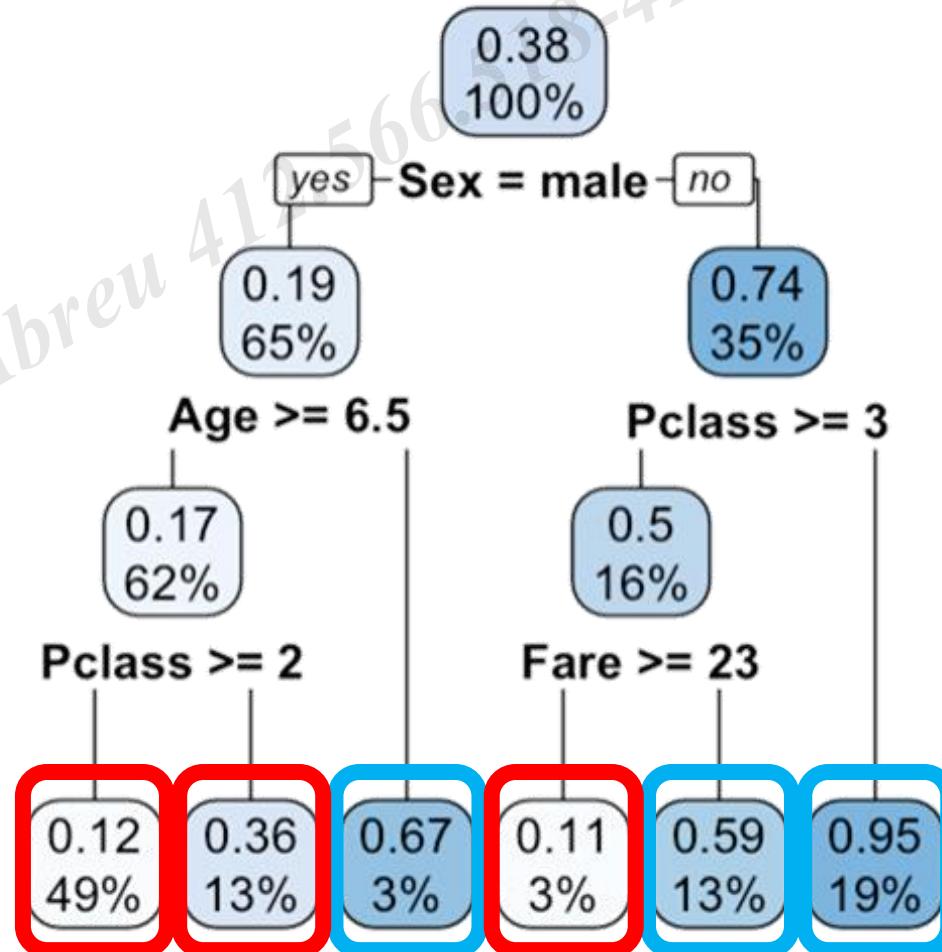
Classificação:

Classificação padrão:

Sobrevivente: $P(S|F) \geq 50\% \Rightarrow C(F) = "Y"$

Não sobreviventes: $P(S|F) < 50\% \Rightarrow C(F) = "N"$

Valor predito	Valor Verdadeiro	
	0	1
0	484	96
1	65	246



Avaliação do modelo



- Acurácia:

Acertos sobre tentativas

Valor predito	Valor Verdadeiro	
0	0	1
1	484	96
	65	246

No exemplo:

$$\frac{484 + 246}{891} = 82\%$$

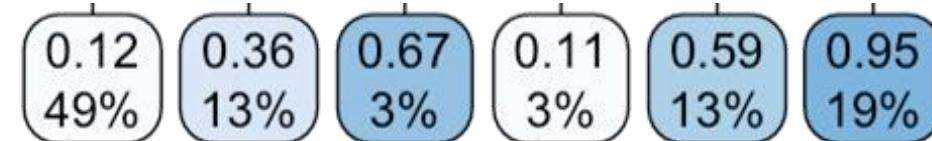
Árvore como diagnóstico

$$\text{Sensitividade: } \frac{TP}{FN+TP} = \frac{246}{246+96} = 72\%$$

$$\text{Especificidade: } \frac{TN}{TN+FP} = \frac{484}{484+65} = 88\%$$

Valor predito	Valor Verdadeiro	
	0	1
0	484	96
1	65	246

Valor predito	Valor Verdadeiro	
	0	1
0	TN	FN
1	FP	TP



Diagnóstico e pontos de corte

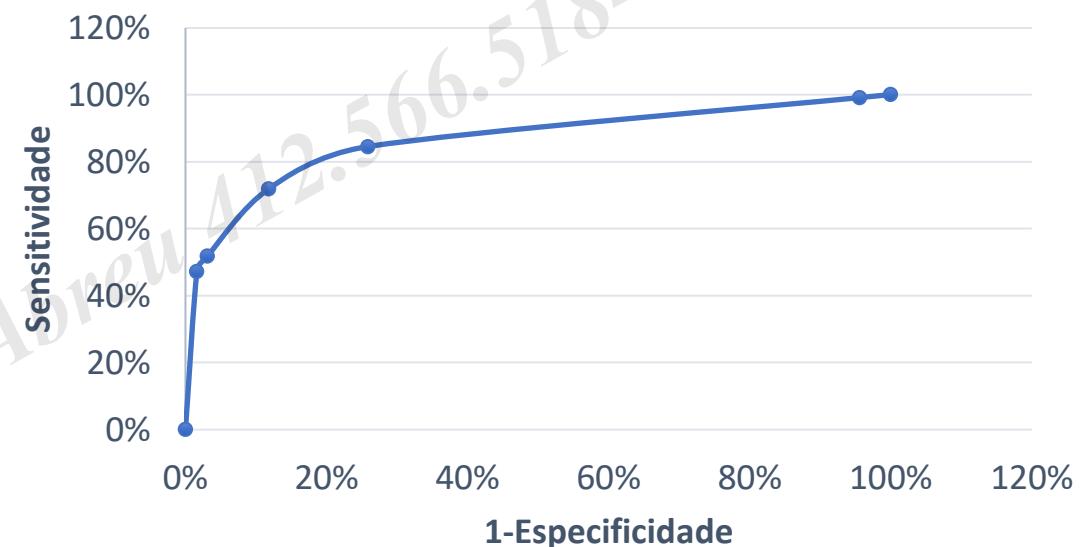
Corte	TP	FP	TN	FN
0% - 11,1%	342	549	0	0
11,1% - 11,5%	339	525	24	3
11,5% - 35,8%	289	142	407	53
35,8% - 58,9%	246	65	484	96
58,9% - 66,7%	177	17	532	165
66,7% - 94,7%	161	9	540	181
94,7% - 100%	0	0	549	342

Acurácia	Especificidade	1-Especificidade	Sensibilidade
38%	0%	100%	100%
41%	4%	96%	99%
78%	74%	26%	85%
82%	88%	12%	72%
80%	97%	3%	52%
79%	98%	2%	47%
62%	100%	0%	0%

Para cada ponto de corte, temos uma matriz de confusão.
No caso, temos 8 possíveis matrizes com a árvore treinada.

Curva ROC

Corte	1-Especificidade	Sensibilidade
0% - 11,1%	100%	100%
11,1% - 11,5%	96%	99%
11,5% - 35,8%	26%	85%
35,8% - 58,9%	12%	72%
58,9% - 66,7%	3%	52%
66,7% - 94,7%	2%	47%
94,7% - 100%	0%	0%



A curva ROC é um gráfico de dispersão de 1-Especificidade no eixo x por Sensibilidade no eixo y, obtidos para cada possível ponto de corte do classificador.



OMML1 _script02-Algoritmo_avaliacao_overfitting

Exercício para casa

script03



OMML1_script03_Exercício para casa



Conclusão

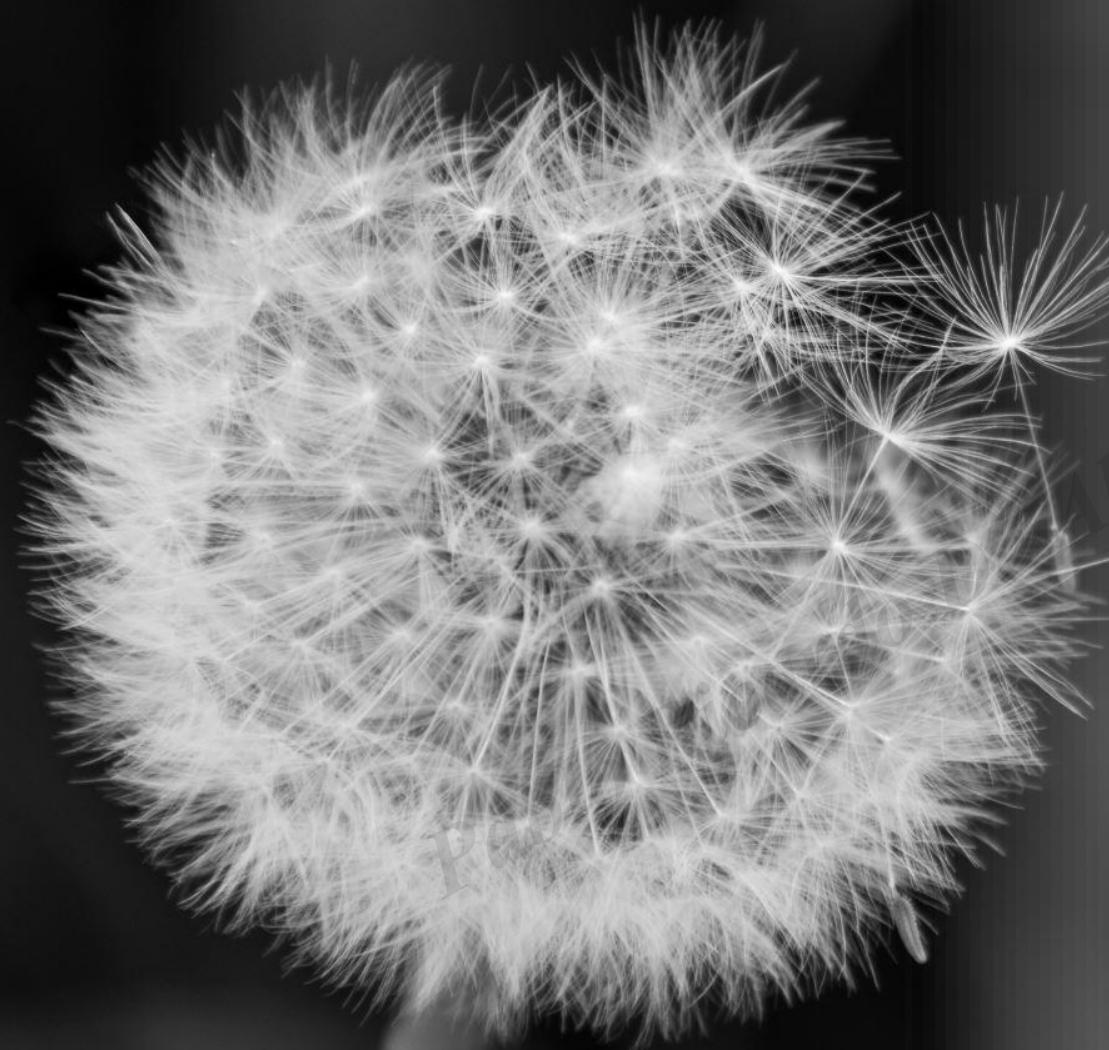
- Robustas, interpretáveis, flexíveis
- Sem suposições probabilísticas
- Necessário *cross-validation*

Quanto mais aprendo, mais
tenho certeza de que, o que
sei, é apenas uma gota,
diante do oceano do que
ainda preciso aprender.



PENSADOR

Jose Ap Barcelos



Por hoje é só ;)



linkedin.com/in/joao-serrajordia

Algoritmos famosos

- CART
- CHAID
- ID3
- C4.5
- C5.0

Stack overflow interessante sobre isso:

<https://stackoverflow.com/questions/9979461/different-decision-tree-algorithms-with-comparison-of-complexity-or-performance>

Na próxima aula...

Técnicas de validação cruzada

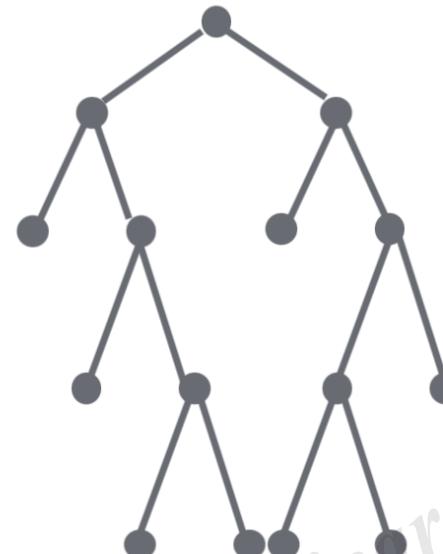
Extras





OMML1 _script02-Algoritmo_avaliacao_overfitting

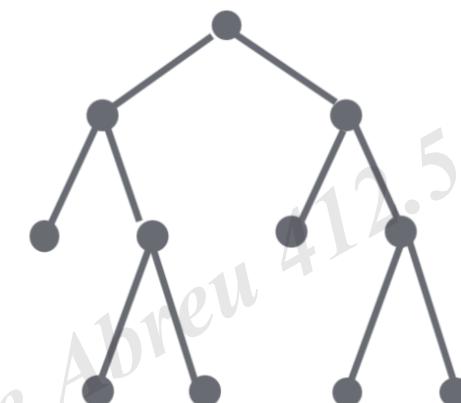
Poda da árvore (*Pruning*)



Acurácia

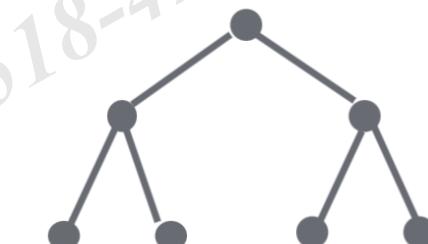
Base de treino: 95%
Base de validação: 40%

Amostra de treino



Base de treino: 70%
Base de validação: 60%

Amostra de validação



Base de treino: 65%
Base de validação: 64%

Estratégias de cross validation

Escolher parâmetros do modelo com uma base de validação ainda pode propiciar overfitting.

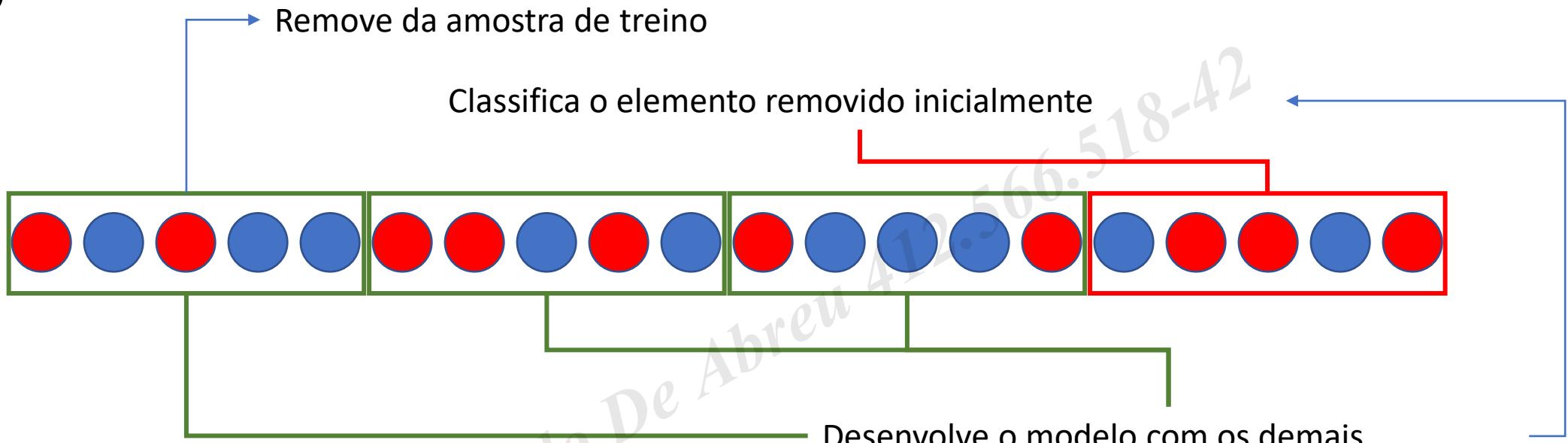
Há diversas técnicas de validação cruzada para se evitar esse efeito. No momento vou mencionar uma técnica clássica: dividir a base em Treino, Validação e Teste

Amostra de treino

Amostra de validação

Amostra de teste

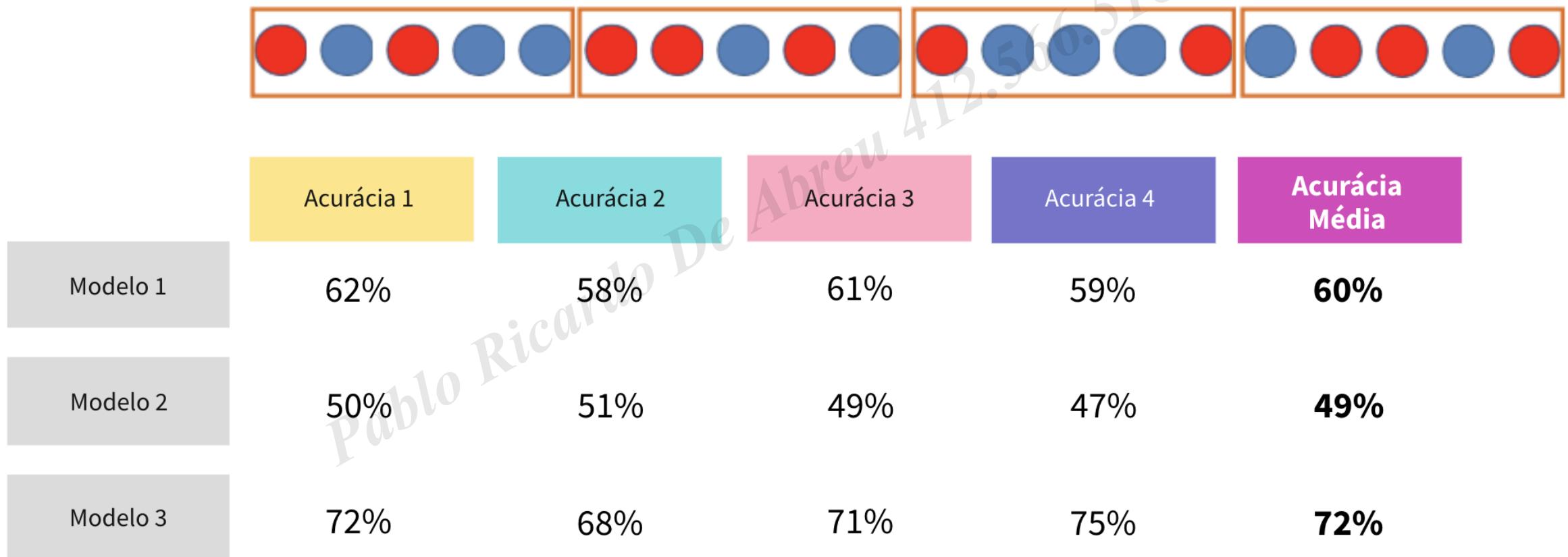
K-fold



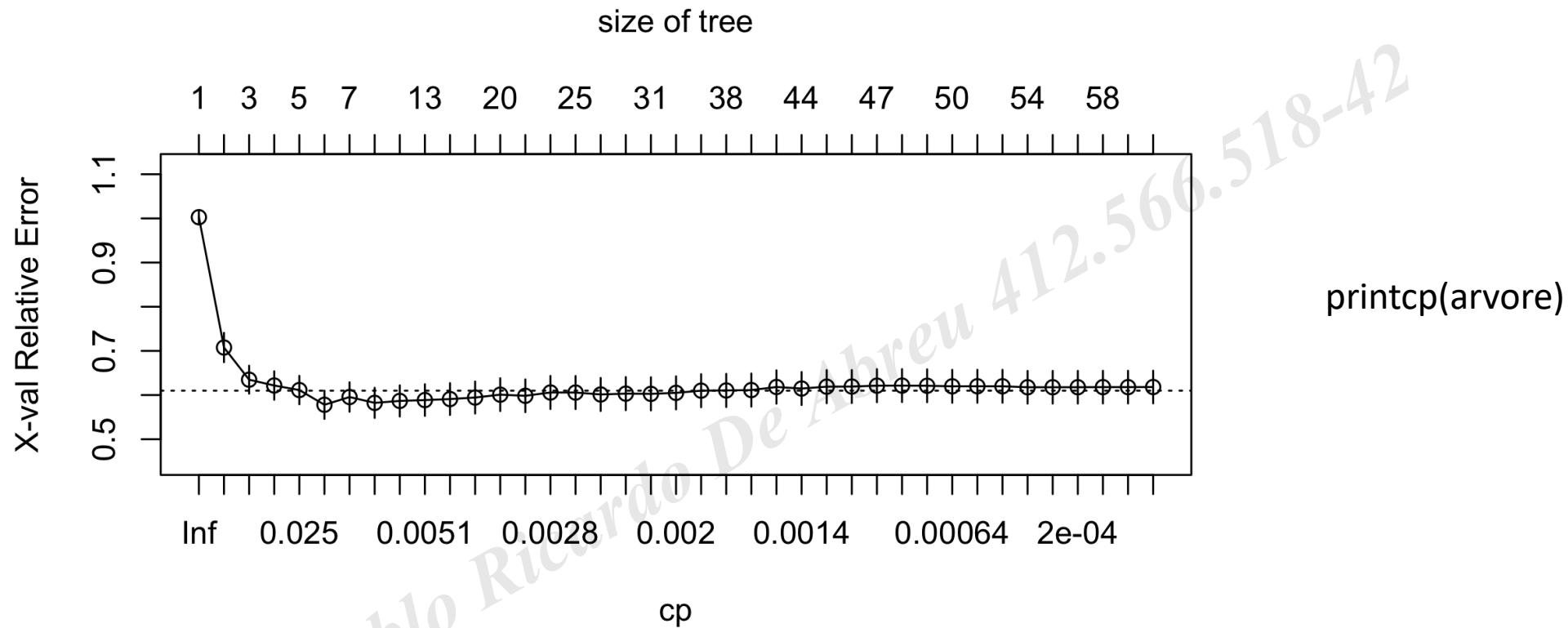
- Dividimos a base em k sub-amostras
- Para cada sub-amostra:
 - Removemos a sub-amostra como validação
 - Treinamos o modelo com as observações restantes
 - Utilizamos este modelo para classificar a sub-amostra removida
 - Avaliamos a métrica de desempenho do modelo
- Calculamos a média das métricas de desempenho do modelo

K-fold

Tipicamente, fazemos o mesmo para variações do modelo para otimizar hiperparâmetros.



Post-prunning com crossvalidation



O R faz a poda da árvore realizando um k-fold para otimizar o CP (complexity path), um parâmetro que sumariza a complexidade da árvore. Isso é feito com um *k-fold*.

OBRIGADO!

linkedin.com/in/joao-serrajordia/