MBA USP ESALQ

ÁRVORES, REDES E ENSEMBLE MODELS

Professor João Fernando Serrajordia Rocha de Mello

Amilcar Ferros Fanelli 708.208-58

MBA USP ESALQ

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

Proibida a reprodução, total ou parcial, sem autorização.

Amilcar Ferraz Faina 188.208.208-58

Lei nº 9610/98

Apresentação

João Fernando Serrajordia Rocha de Mello – (Juka)

Amilcar Ferraz Farina 180.708.208-58

Trajetória profissional

Modelagem de crédito em grandes bancos
Desenvolvimento de modelos
Validação de modelos

Professor universitário / MBA / Pós Lato Sensu
Consultoria em ciência de dados

Acadêmico

IME USP

BACHAREL EM ESTATÍSTICA

MESTRE EM ESTATÍSTICA

*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

O que vamos aprender?

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

Amilcar Perraz Fa 8 308-58

*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

Agenda de hoje

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

Amilcar Ferraz Fa 8 208-58

*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

Vamos discutir...

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

188.708.208-58

*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

O que é Machine Learning?

Marraz Farina 108.708.208-58

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

*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

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
rec

"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."

Anibal Ferraz Farina 188.708.208-58

*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

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/

Amilcar Ferraz Farias 108.708.208-55

*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

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/

Annicar Perraz Farina 108.505.208-58

*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

Vamos discutir...

E o que é aprendizado de máquina...?

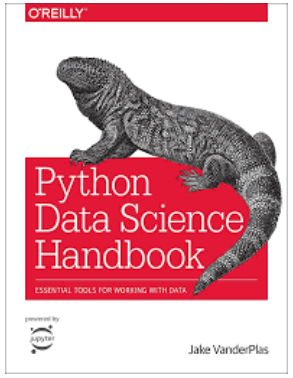
188.708.208-58

*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

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



Amilcar Ferraz Farina 188.708.208-58

*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

Preparativos

- Abrir o Spyder
- Instalar os pacotes
  - ARNEM_aula1_script00_pacotes.py
- Planilha eletrônica
- Algo para fazer suas anotações

Emilio Velez Farina 188.708.248-58

*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

Árvores de decisão

Amilcar

*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

Árvores de decisão:

Onde vivem?

O que são?

O que comem?

Predadores naturais

Andrea Pereira Farina 188.708.208-58

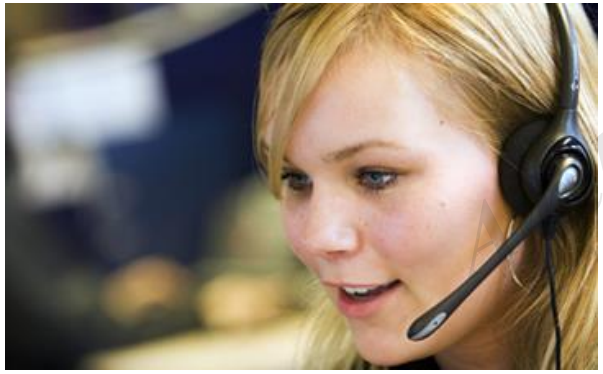
*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.
Proibida a reprodução, total ou parcial, sem autorização. Lei nº 9610/98

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?

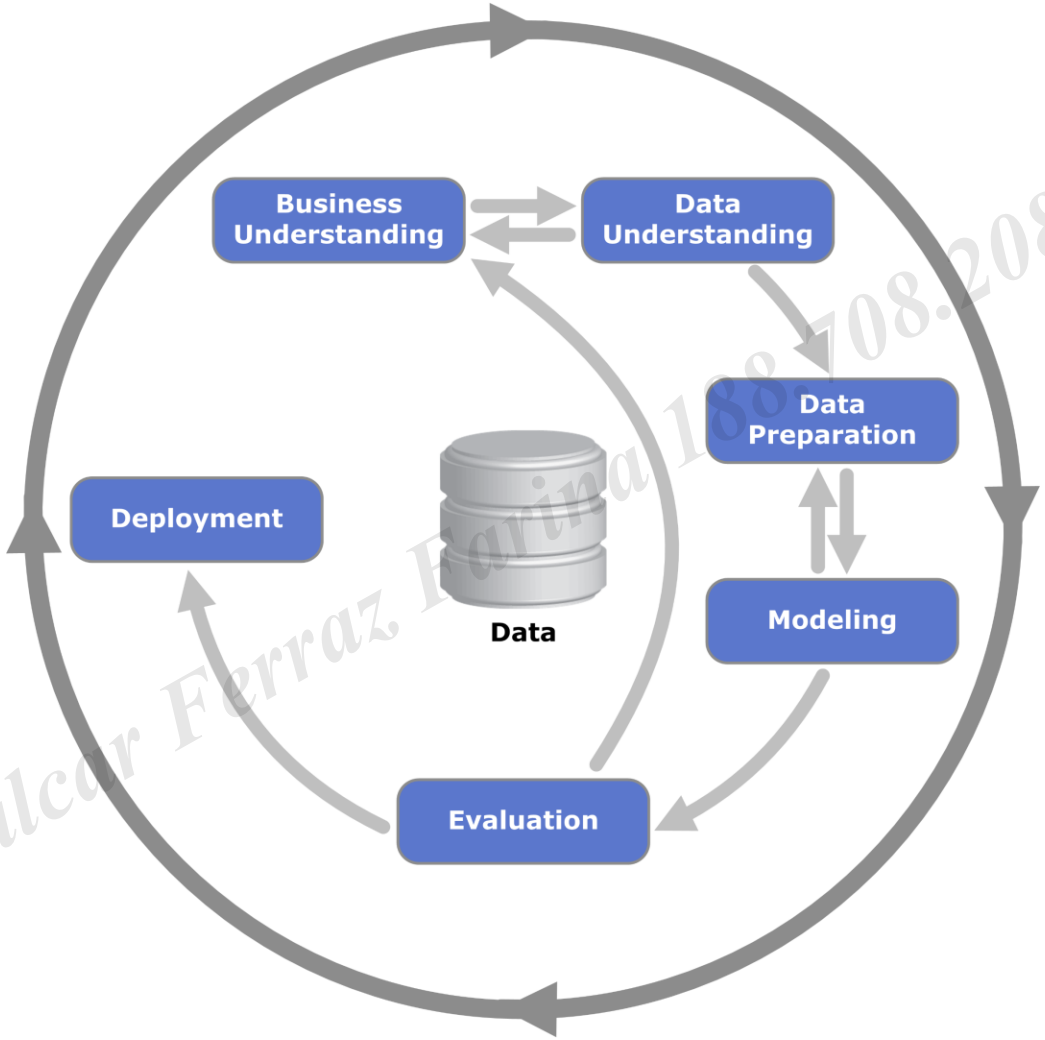
O que a pessoa está fazendo?

Quão ecológico esse veículo é?

Alcar Ferraz Farina 188.708.208-58

*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

CRISP-DM



Amilcar Ferraz Farina 188.708.208-58

Fonte: https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png

*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

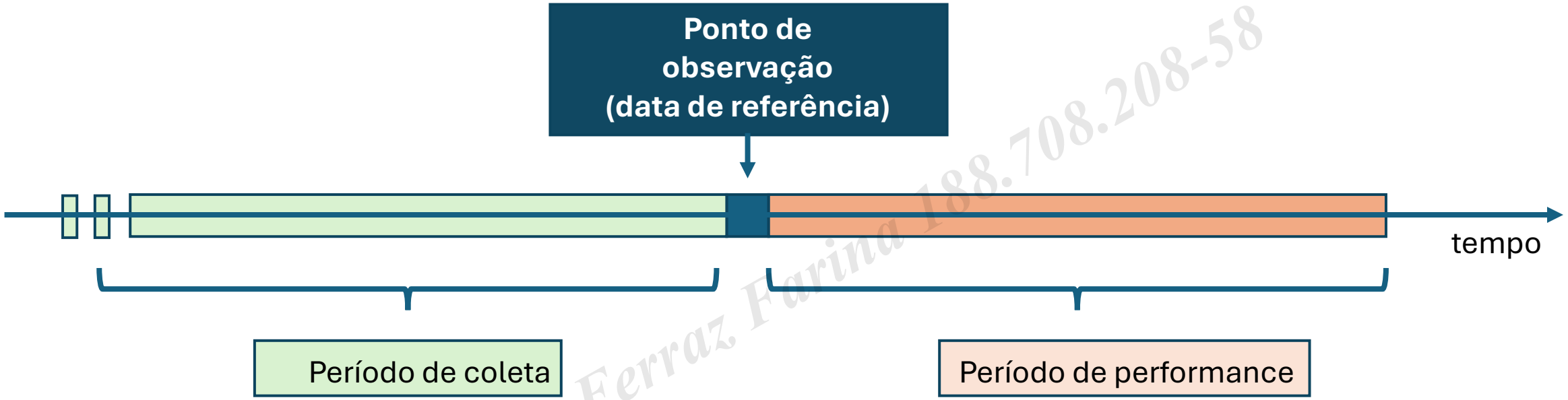
Modelos preditivos

Como é isso?

Amilcar Hanaz Farina 188.708.208-58

*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

Desenho de safra (ou coorte)



Amilcar Ferraz Farina 188.708.208-58

Exemplo de desenho amostral para modelo preditivo

*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

Estrutura da base de dados

Cada coluna é uma variável

Cada linha representa uma observação

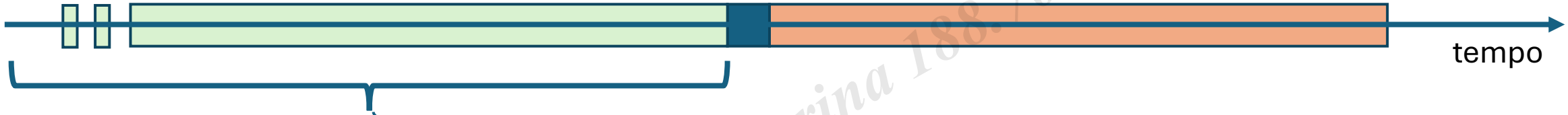
ID_cliente	Idade	Renda	Tempo de emprego	Negativado_Bureau	...	Inadimplente_em_12m
2914901	60	75579	10	0	...	0
5003858	25	33495	2	0	...	1
8309347	51	59908	9	0	...	1
4637814	46	53470	10	0	...	0
7203409	65	69822	1	0	...	1

Amilcar Terraz Farina 188.706.208-58

*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

Desenho amostral

Informações disponíveis ANTES da decisão.

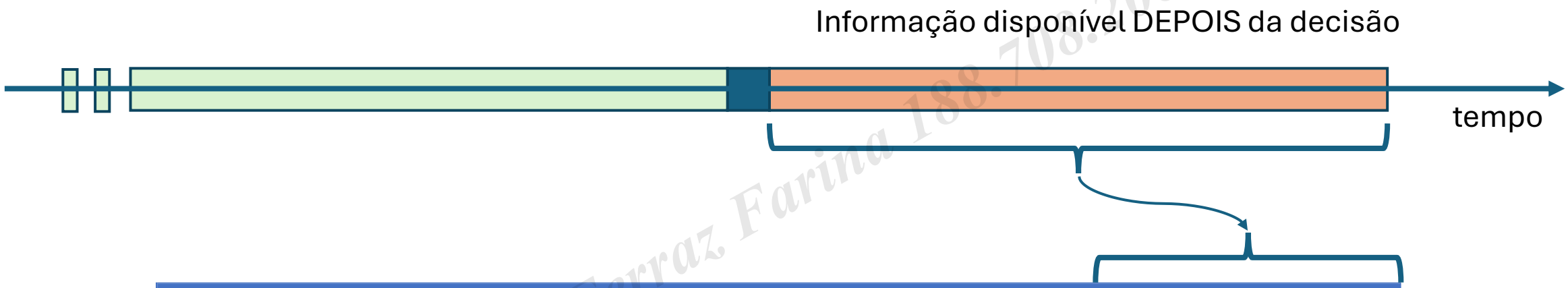


ID_cliente	Idade	Renda	Tempo de emprego	Negativado_Bureau	...	Inadimplente_em_12m
2914901	60	75579	10	0	...	0
5003858	25	33495	2	0	...	1
8309347	51	59908	9	0	...	1
4637814	46	53470	10	0	...	0
7203409	65	69822	1	0	...	1

Anibal Ferraz Farina 188.708.208-58

*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

Desenho amostral



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

Anibal Ferraz Farina 188.708.208-58

*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

Previsão do modelo - *treinamento*

$$PD = F(X1, X2, \dots Xn)$$

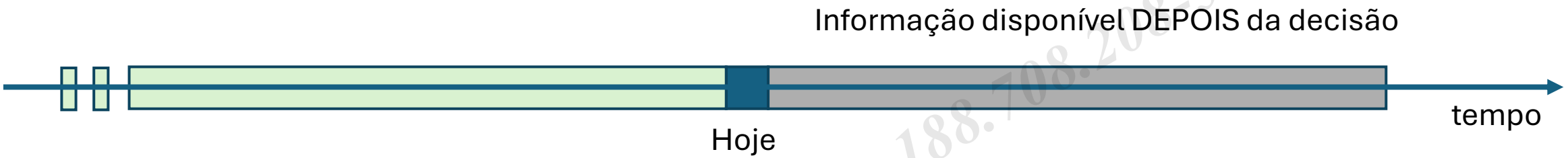
Previsão – probabilidade de inadimplência

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%

Amilcar Fernez Farina 188.708.208-58

*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

Previsão do modelo - *aplicação*



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

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

Amilcar Ferraz Farma 188.708.208-58

*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

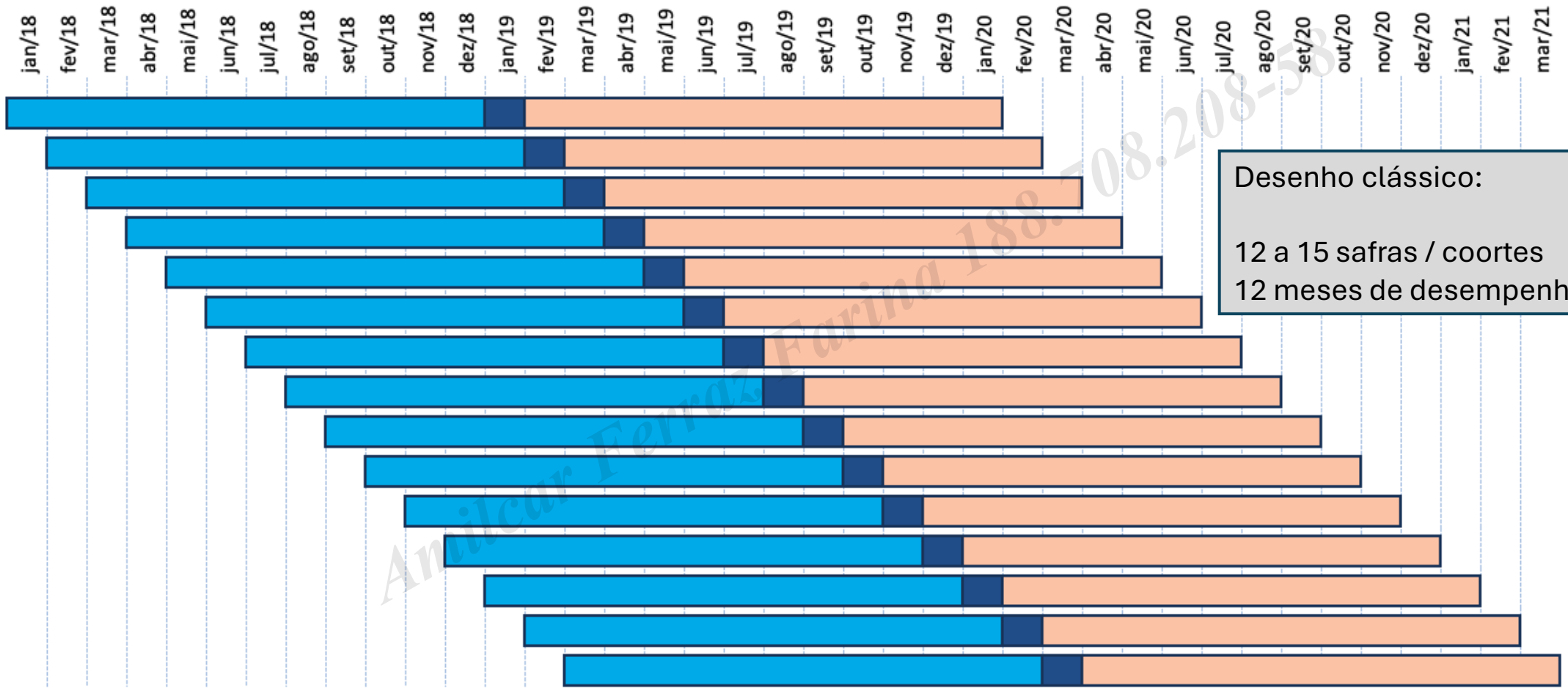
Desenho clássico

- 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”)

Amilcar Ferraz Farina 108.708.208-58

*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

Desenho do modelo



Desenho clássico:

12 a 15 safras / coortes
12 meses de desempenho

Anilcar Ferraz Farina 188.708.208-58

*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

Árvores de decisão:

Onde vivem?

O que são?

O que comem?

Predadores naturais

Amilcar Carina 188.708.208-58

*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

Classificação dos algoritmos

Paradigma Machinelârnico

- 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** (Testes, intervalos de confiança etc)

Estamos aqui!

Amilcar Ferraz Fa 88.7 208-58

*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

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!

Amilcar Ferraz Fa 88.702.208-58

*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

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

Amilcar Ferraz Farina 188.708.208-58

Estamos aqui!

*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

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!

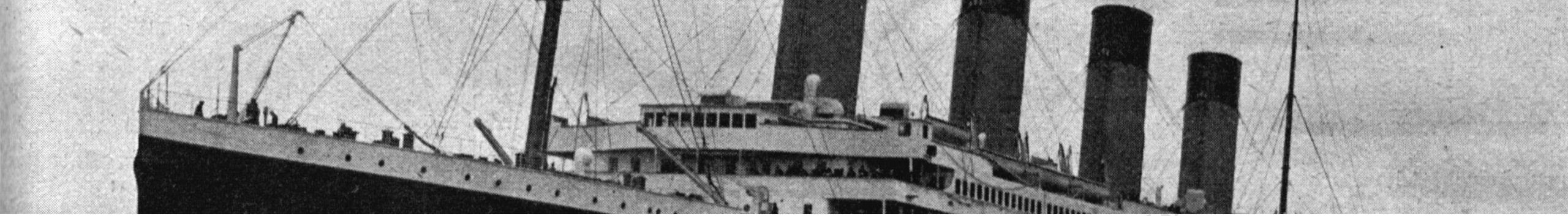
Amilcar Ferraz Farina 188.708.098-58

*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

Nosso problema: classificar sobreviventes

Imagem: 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

Amilcar Ferraz Farina 188.708.208-58

População

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

Amostra

- 891 pessoas
- 549 não sobreviventes
- 342 sobreviventes

*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

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

Anibal Ferraz Farina 188.***.***-20**

*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

Amilcar Ferraz Farias 188.108.208-58

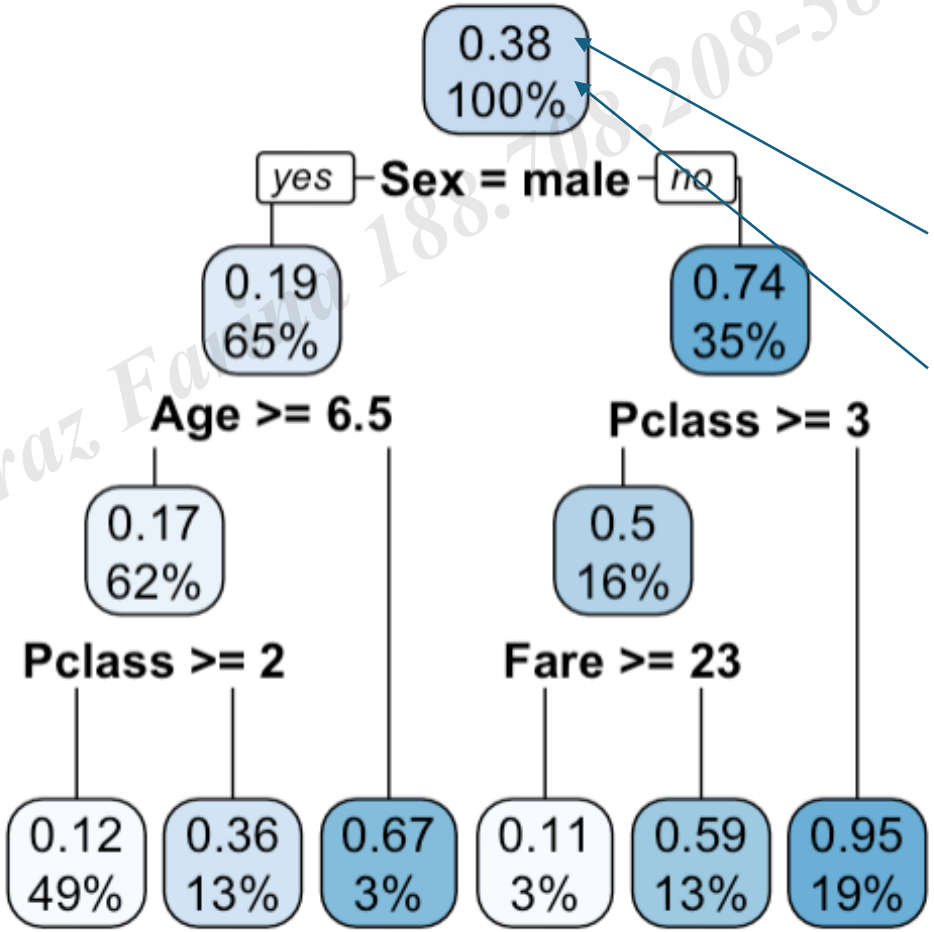
ARNEM_script01-Primeiro_contato_com_arvores

*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

O que é uma árvore de decisão?

A árvore de decisão é:

Uma sequência de segmentações binárias
Que visa homogeneidade da variável
resposta



Taxa de sobreviventes

% da amostra

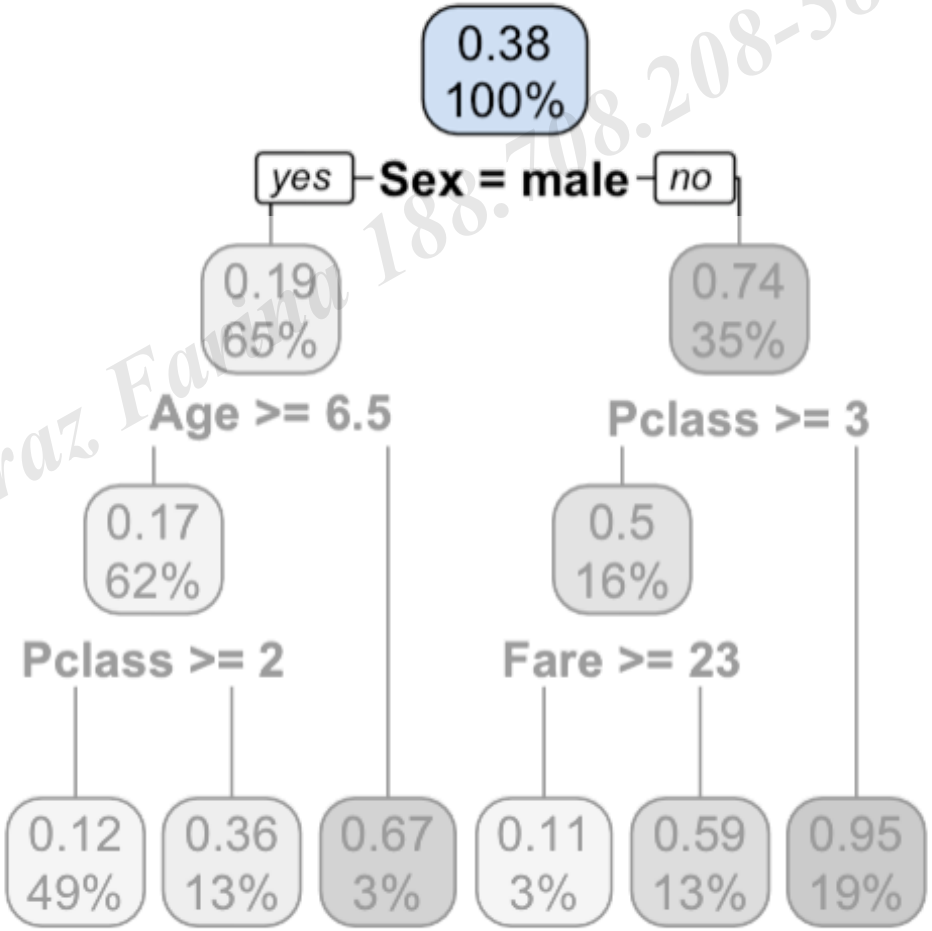
Amilcar Ferraz Farma 188.708.208-58

*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

O que é uma árvore de decisão?

Inicialmente temos 891 passageiros dos quais

342 sobreviveram (38%)
549 não sobreviveram



Amilcar Ferraz Farina 188.708.208-58

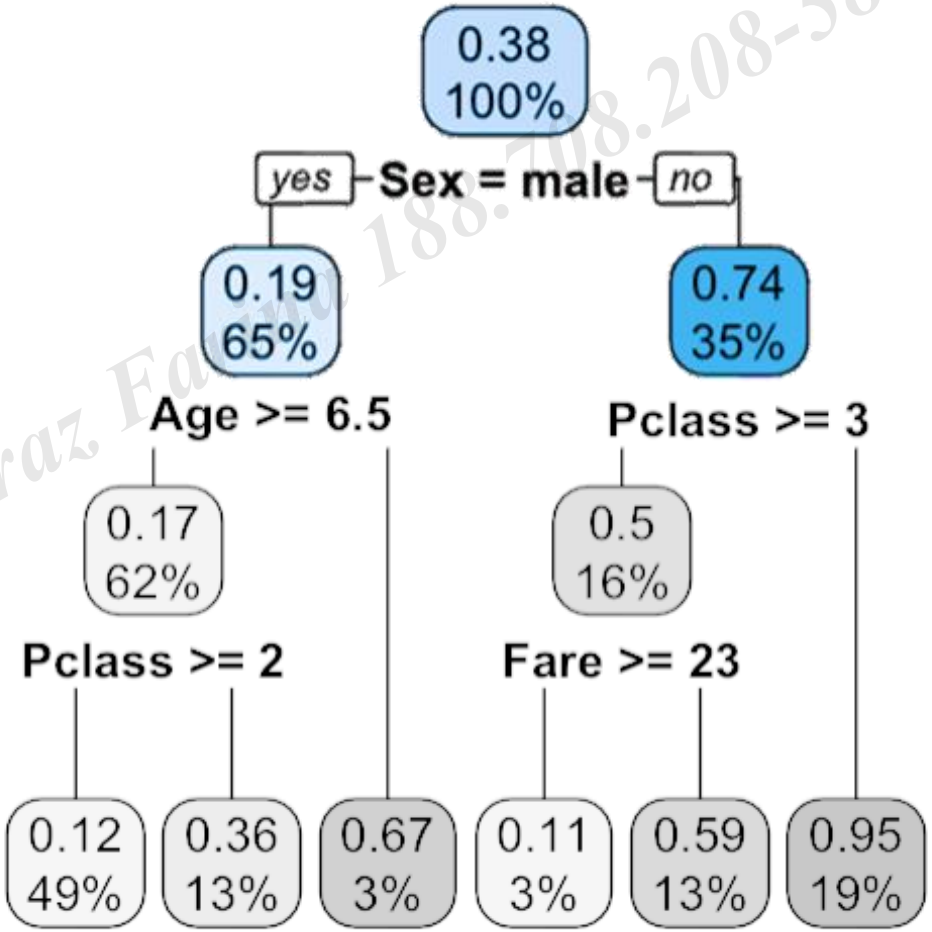
*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.
Proibida a reprodução, total ou parcial, sem autorização. Lei nº 9610/98

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



Amilcar Ferraz Farma 188.708.208-58

*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

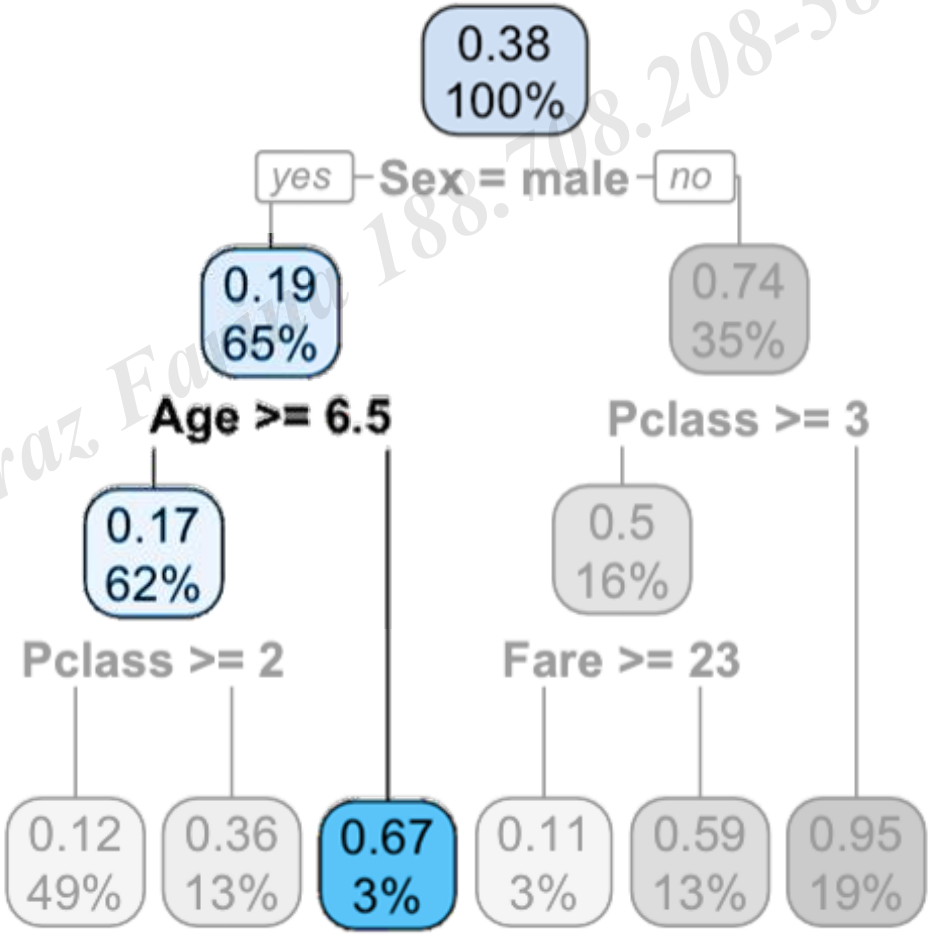
O que é uma árvore de decisão?

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 (>=6,5 anos) dos quais
93 sobreviveram (17%)
553 não sobreviveram

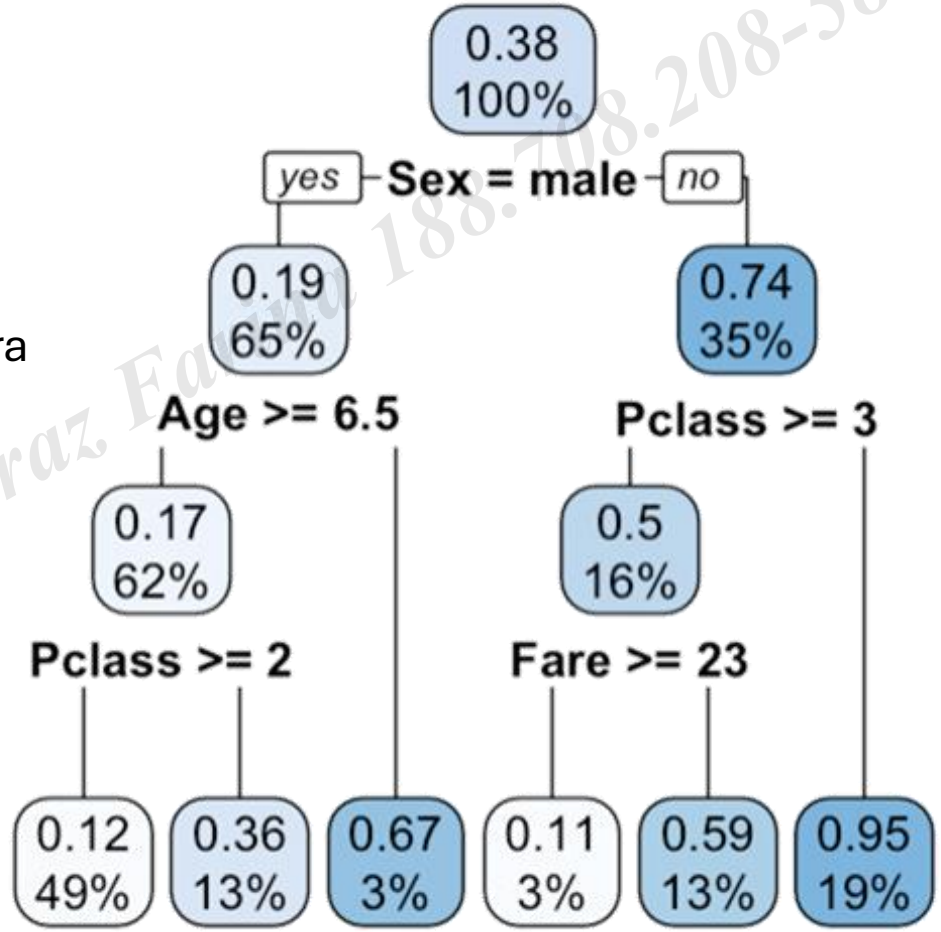


Amilcar Ferraz Farina 188.708.208-58

*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

O que é uma árvore de decisão?

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

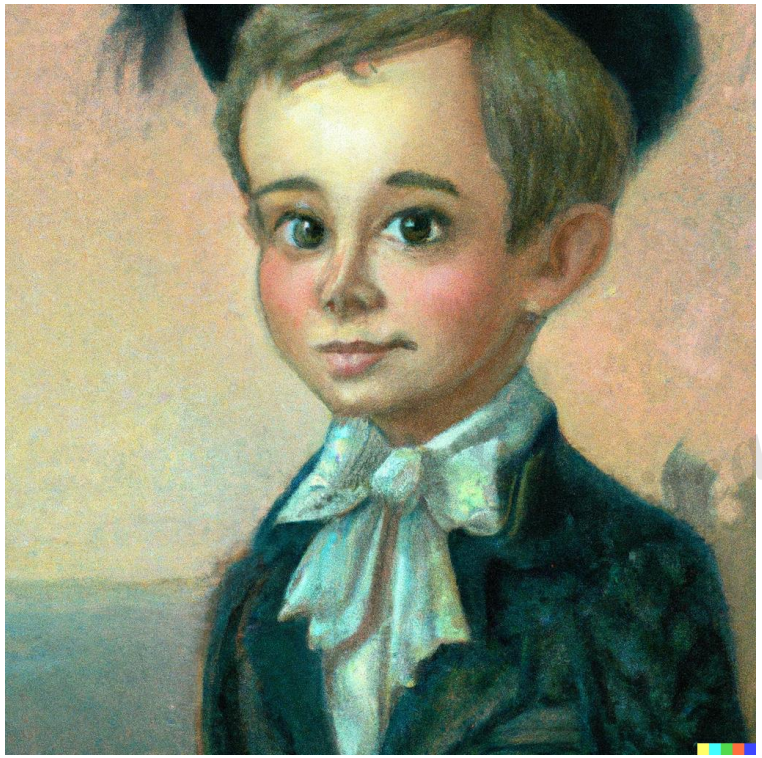


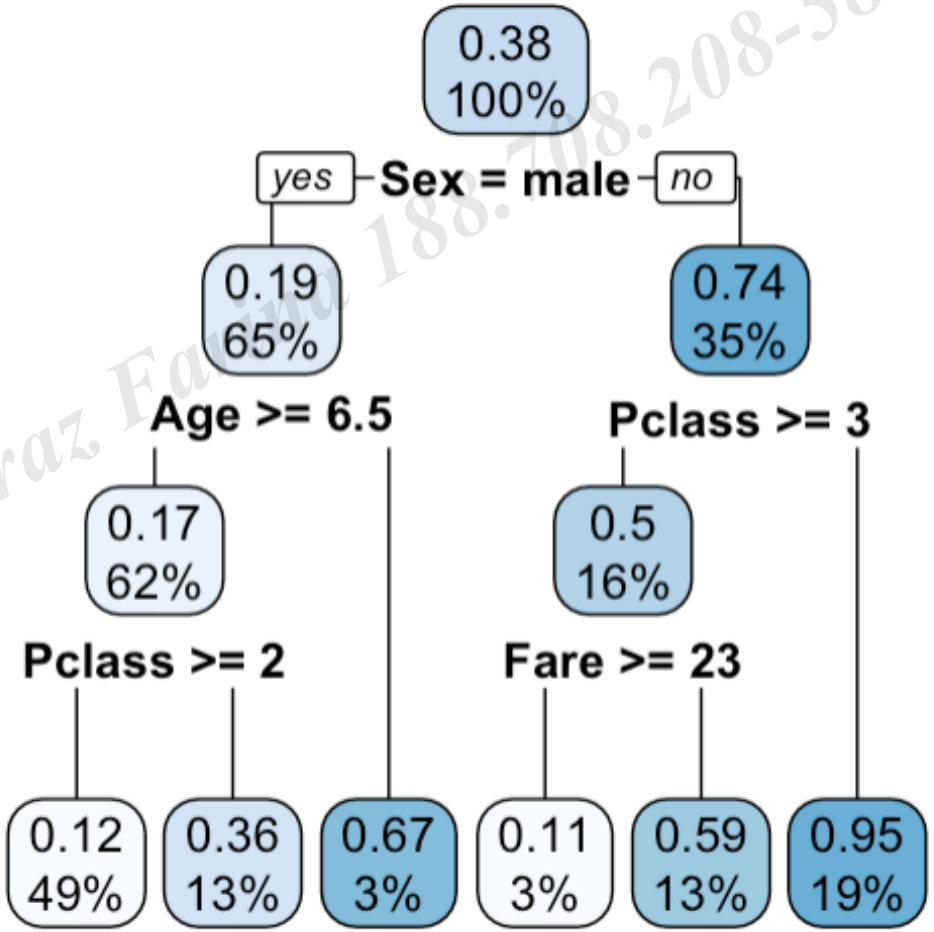
Amilcar Ferraz Farma 188.708.208-58

*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

Vamos experimentar??

Vamos classificar este garoto:



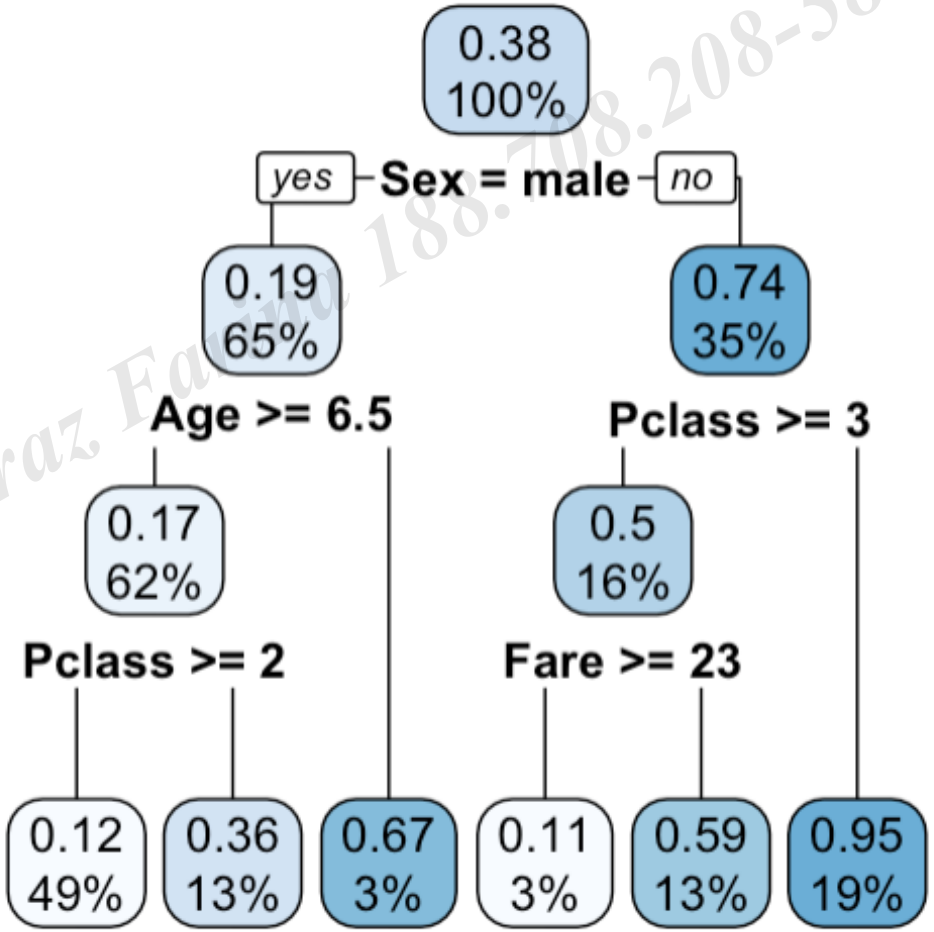


Ferraz Farma 188.708.208-58

*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

Vamos experimentar??

Vamos classificar este pobre homem:

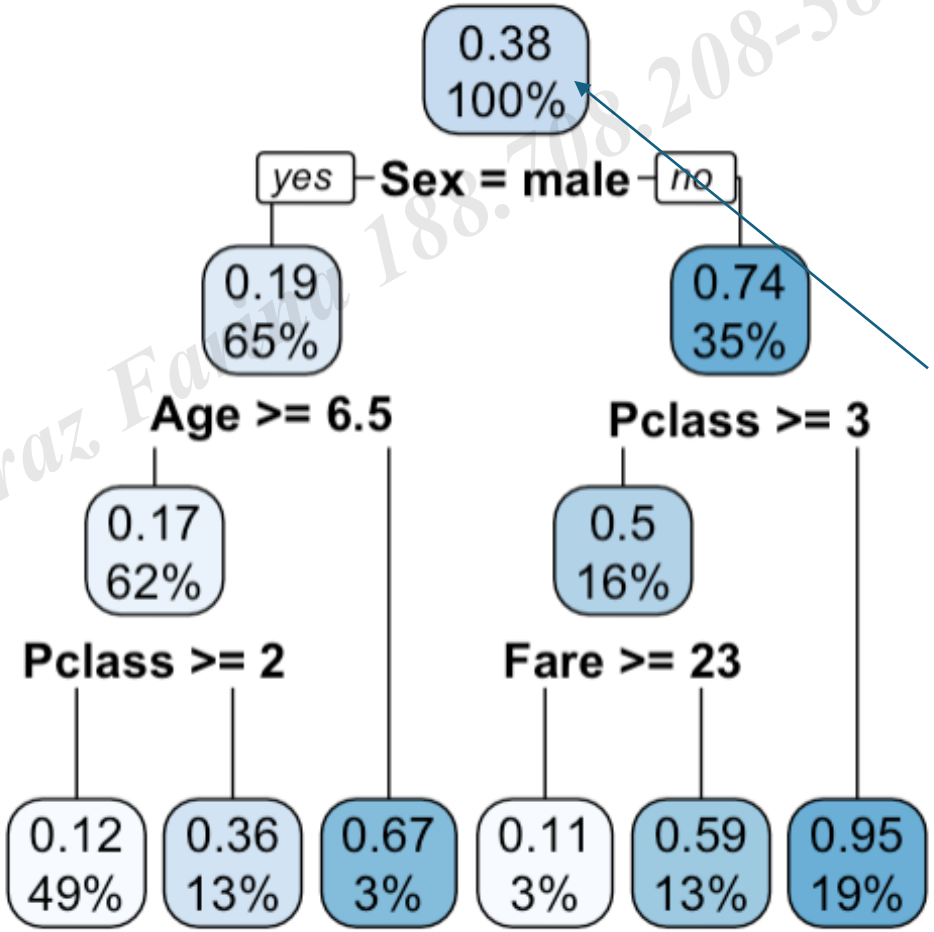


Lucian Ferraz Faria 188.708.208-58

*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

Vamos experimentar??

Vamos classificar este rico senhor:



Taxa de sobreviventes

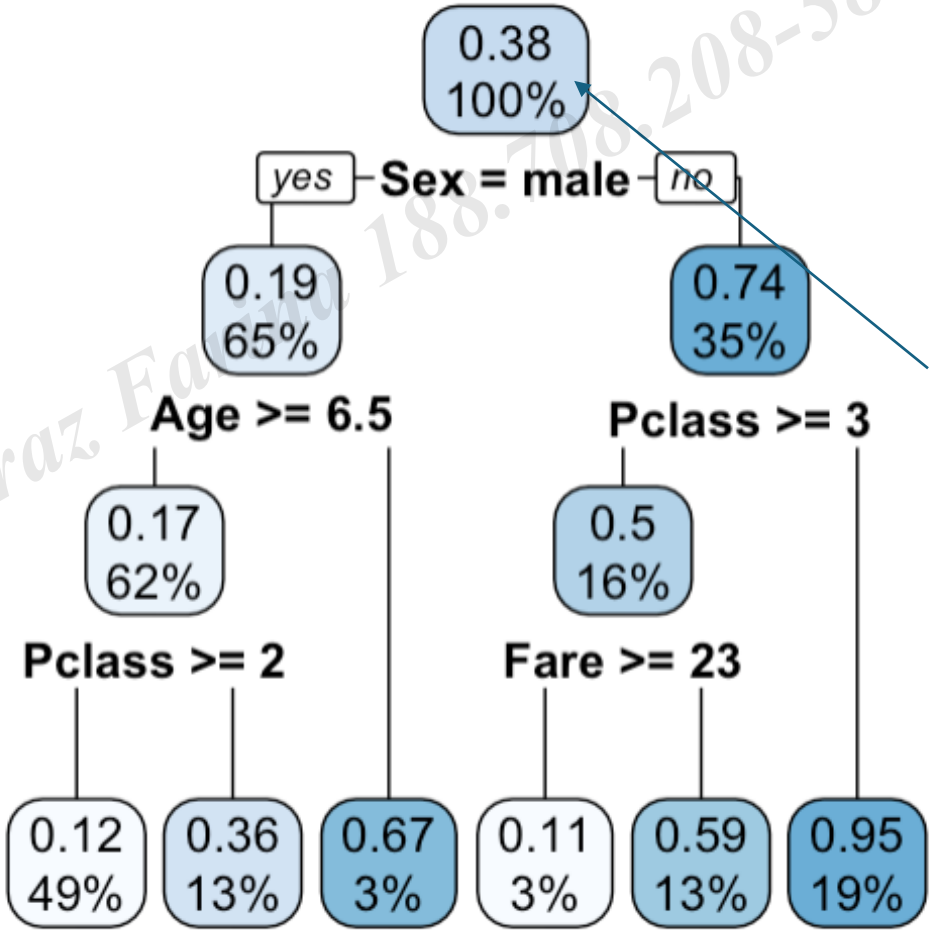
% da amostra

Ana Ferraz Farina 188.708.208-58

*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

Vamos experimentar??

Vamos classificar esta rica senhora:



Taxa de sobreviventes

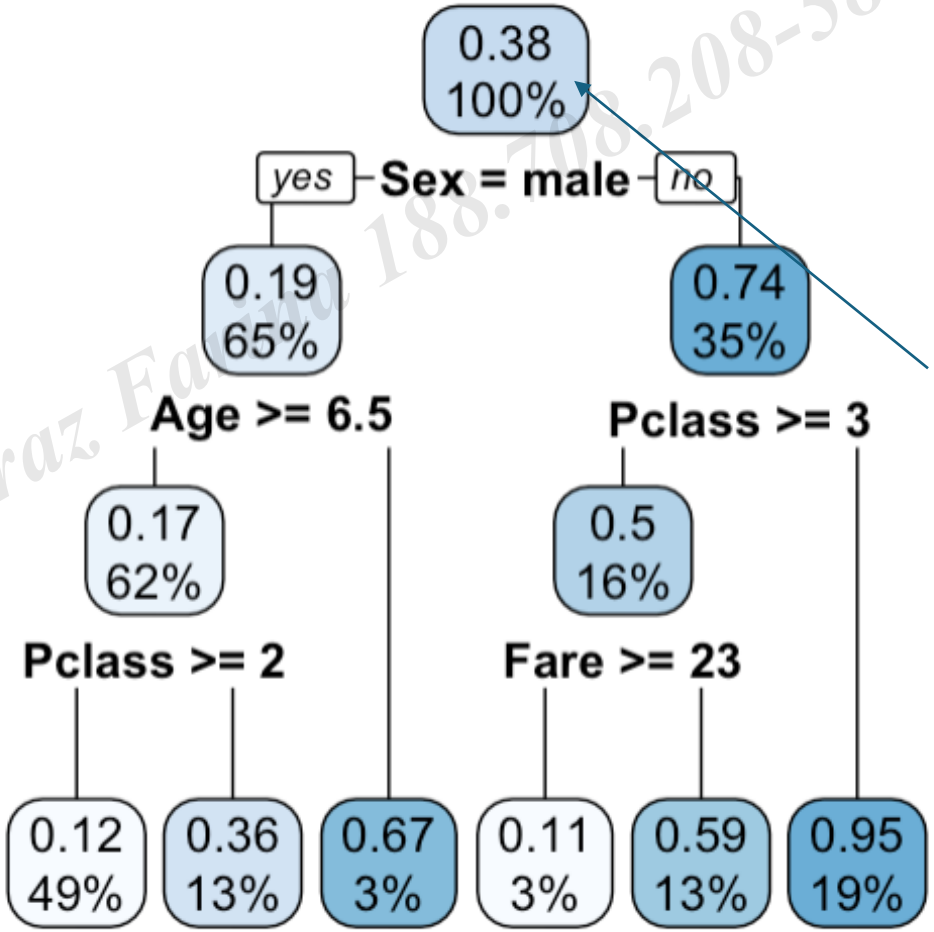
% da amostra

Alexandre Ferraz Farias 188.708.208-58

*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

Vamos experimentar??

Vamos classificar esta pobre moça:



Taxa de sobreviventes

% da amostra

Oscar Ferraz Farma 188.708.208-58

*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

Árvores de decisão:

Onde vivem?

O que são?

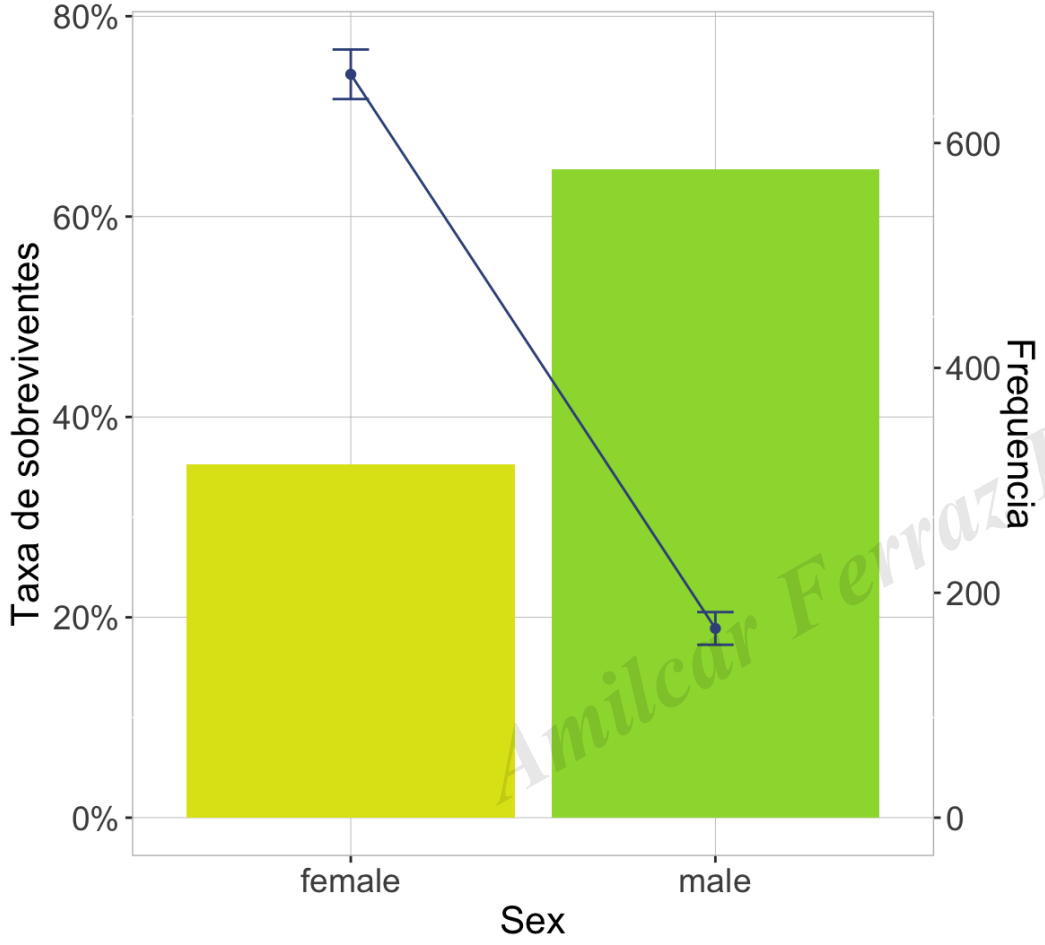
O que comem?

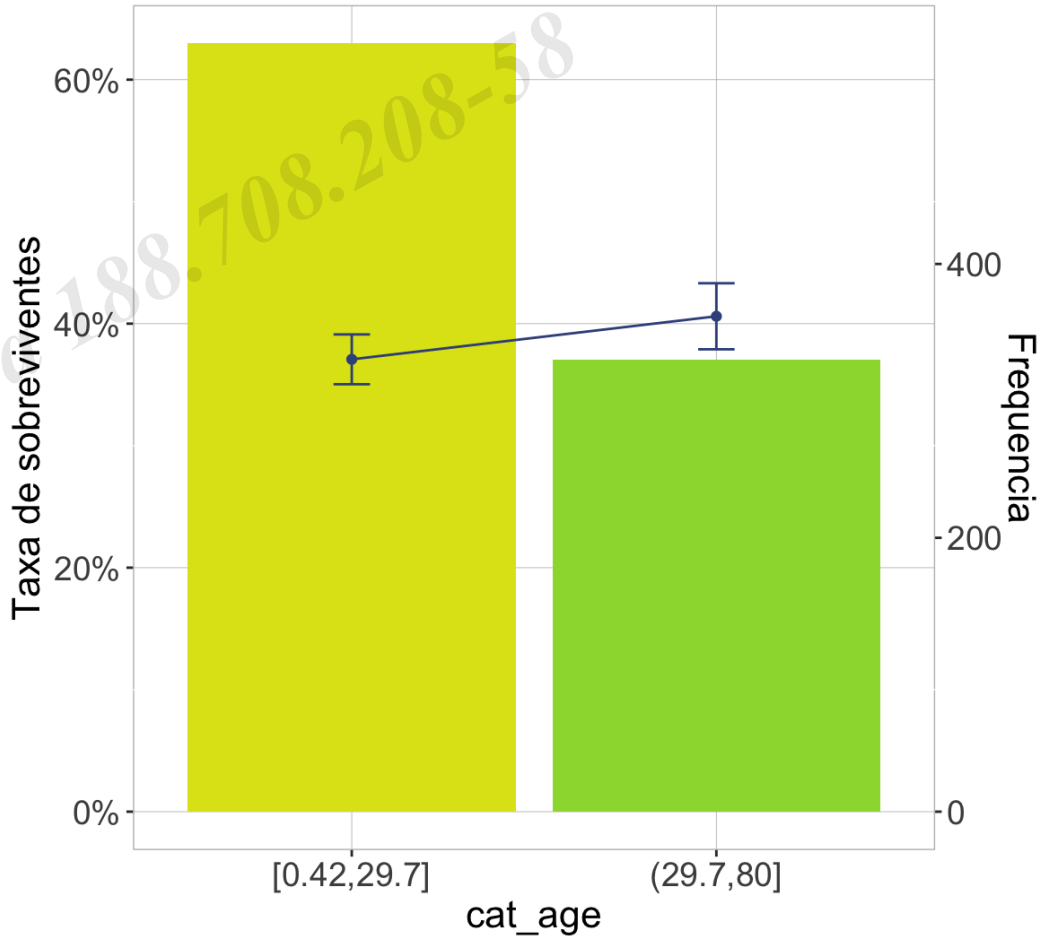
Predadores naturais

Andrea Leonardo Farina 188.708.208-58

*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

Qual é 'melhor'?

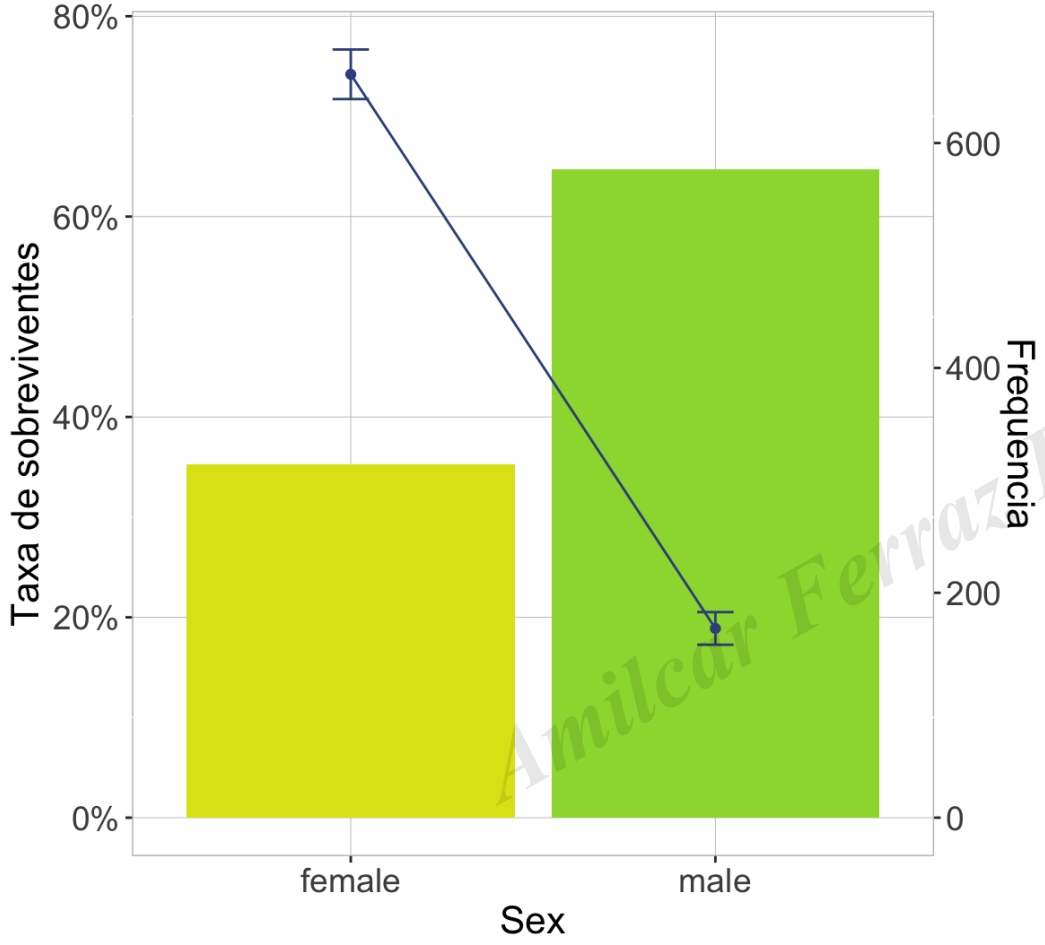


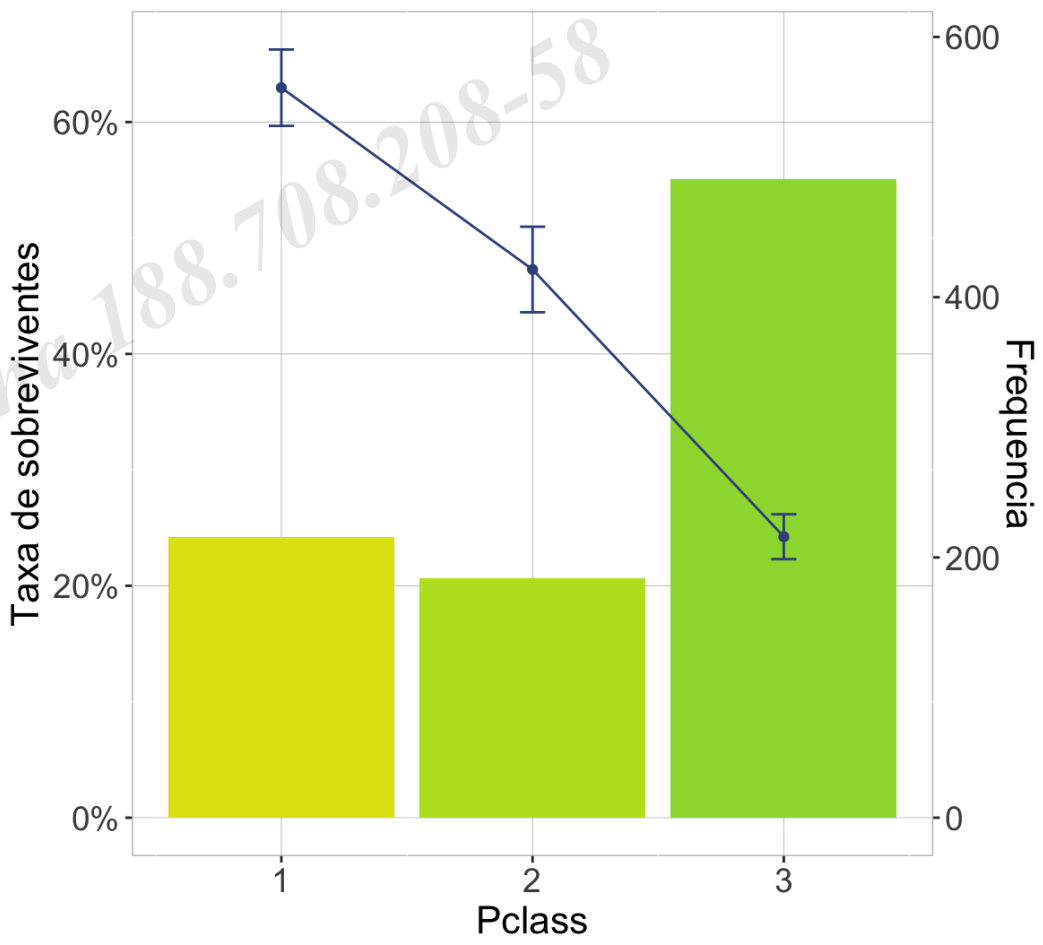


Amilcar Ferraz Farina 188.708.208-58

*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

E agora?





Amilcar Ferraz Farina 188.708.208-58

*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

Medida de “informação”

- Precisamos de uma métrica para ajudar a decidir qual variável agrega 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”:

Amilcar Ferraz Farina 188.708.208-58

*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

Definições de impureza

- Gini
- Entropia de Shannon

Anilcar Ferraz Farina 188.708.208-58

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

*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

Índice de Gini

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

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

Amilcar Ferraz Farina 188.708.208-58

*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

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

Amilcar Ferraz Facina 188.708.208-58

*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

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

Amilcar Ferraz Farina 188.708.208-58

Explicação interativa:

https://view.genially.com/62094a9c11415700197ff277/presentation-algoritmo-de-arvore

*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

Hiperparâmetros

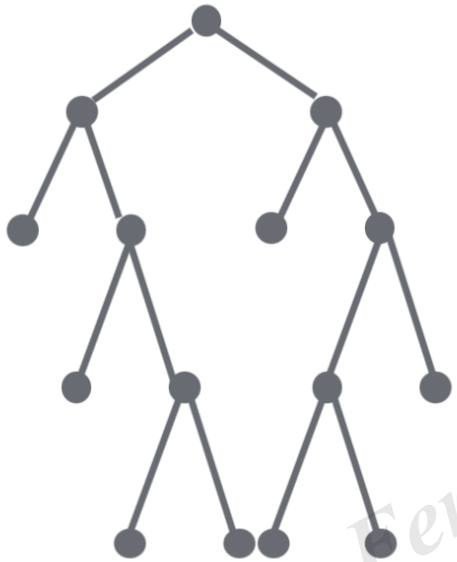
Amilcar Ferraz Fatima 188.708.208-58

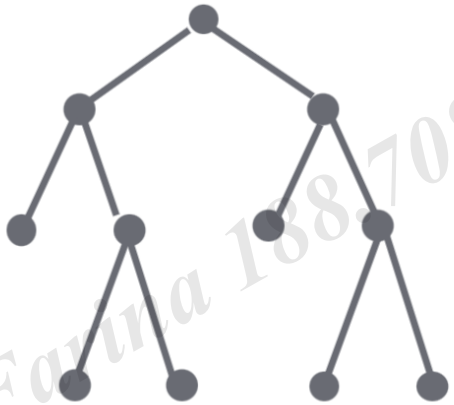
São parâmetros que controlam o algoritmo como:

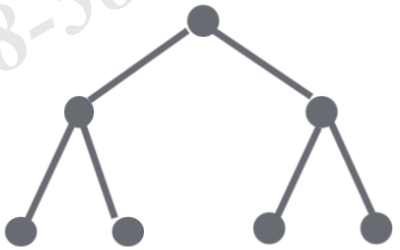
1. Número mínimo de observações por folha
2. Profundidade máxima
3. CP – Custo de complexidade

*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

Custo de complexidade





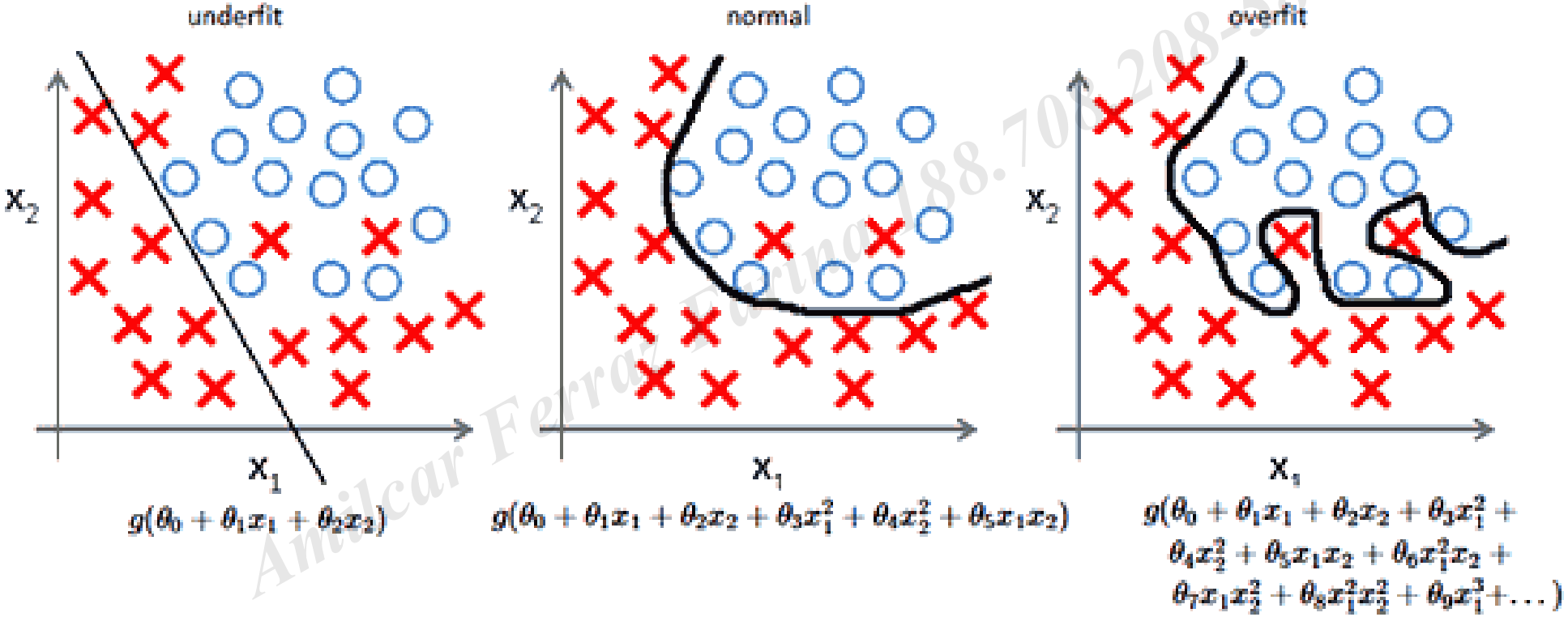


Anilcar Ferraz Farina 188.708.208-58

Custo de Complexidade	Baixo	Médio	Alto
Complexidade resultante	Alta	Média	Baixa

*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

Over fitting / Under fitting

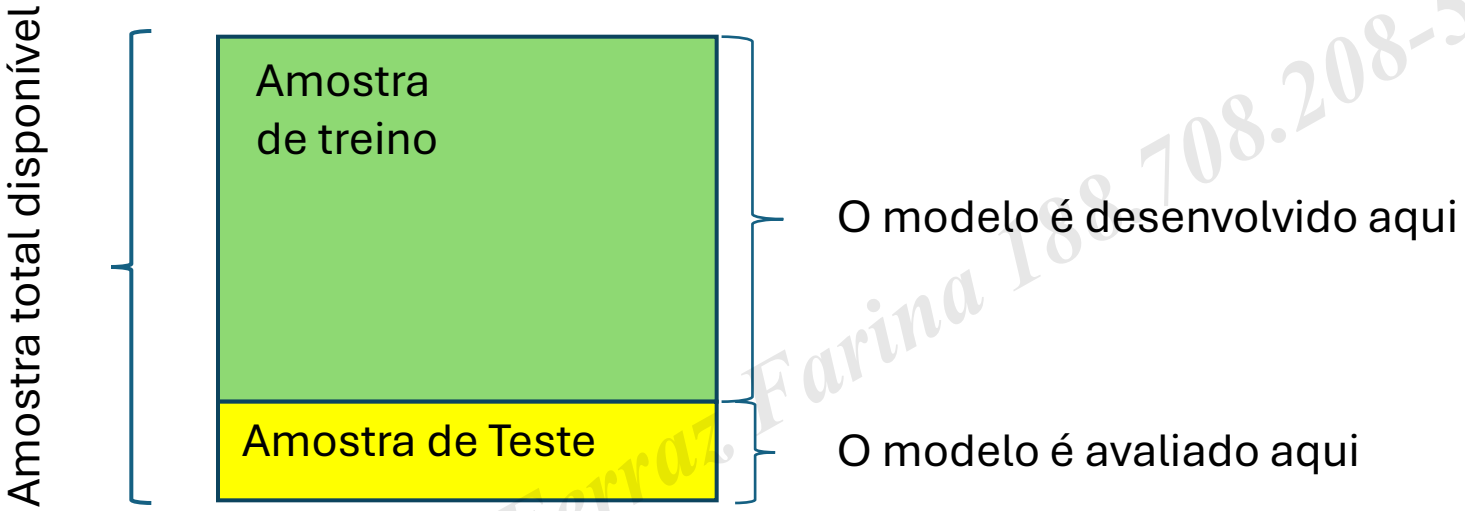


http://mlwiki.org/index.php/Overfitting

Anilcar Ferraz Farina 188.703.208-58

*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

Cross validation



Amilcar Ferraz Farina 188.708.208-58

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.

*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

Amilcar Ferraz Farias 188.108.208-58

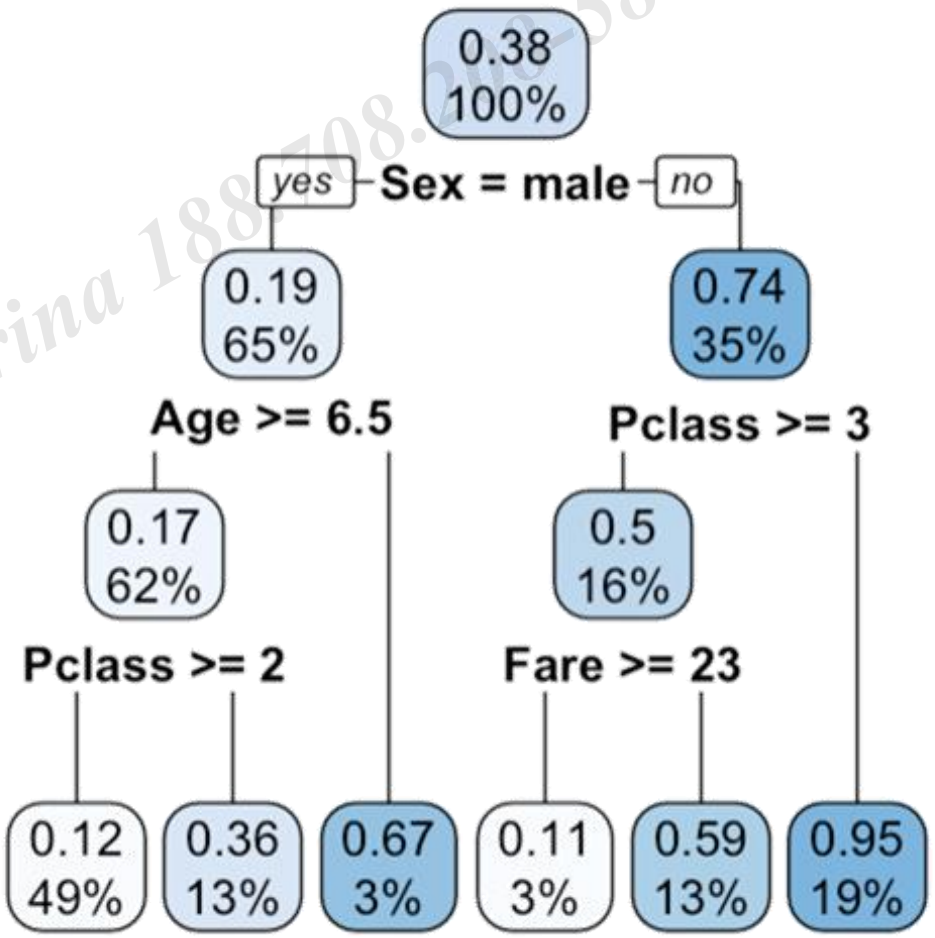
ARNEM _script02-Algoritmo_avaliacao_overfitting

*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

A árvore como um classificador

Requisitos:

Ter todas as variáveis.



Amilcar Ferraz Farina 188.798.288-58

*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

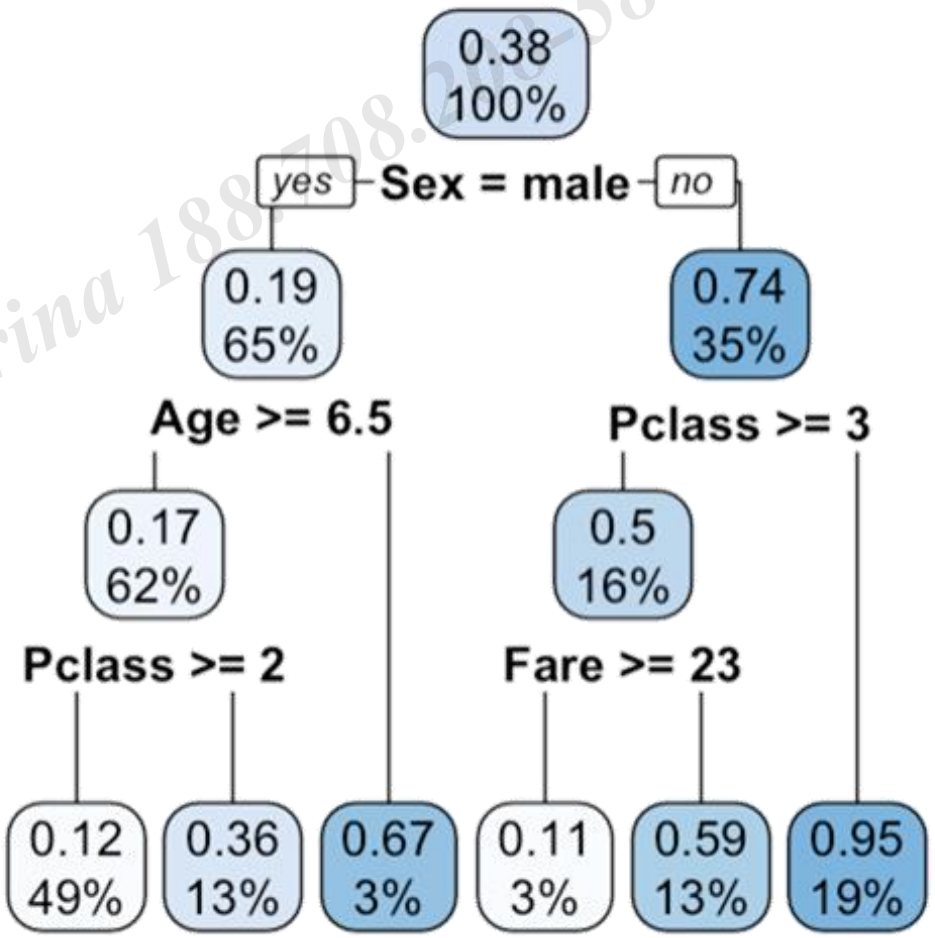
A árvore como um classificador

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



Amilcar Ferraz Farina 188.798.288-58

*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

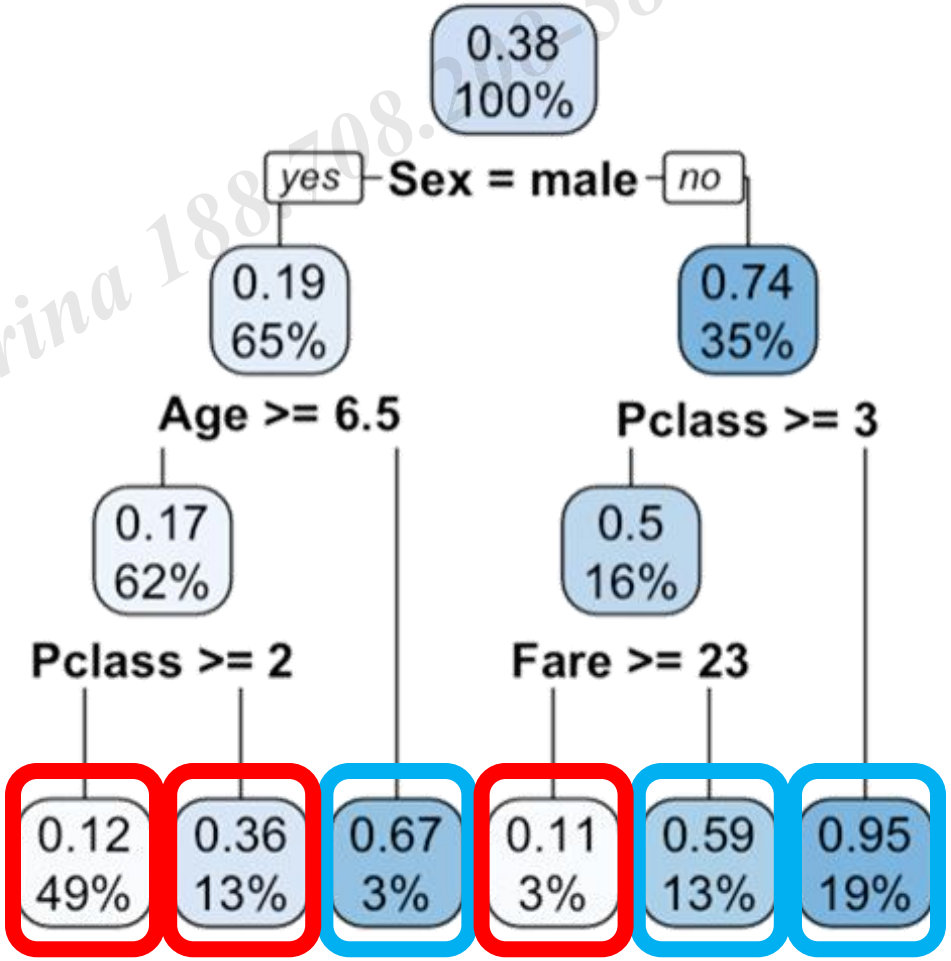
A árvore como um classificador

Classificação:

Classificação padrão:

Sobrevivente: $P(S|F) \geq 50\% \implies C(F) = \text{“}Y\text{”}$
Não sobreviventes: $P(S|F) < 50\% \implies C(F) = \text{“}N\text{”}$

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



Amanda Ferraz Farina 188.798.288-58

*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

Avaliação do modelo

- Acurácia:

Acertos sobre tentativas

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

No exemplo:

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

Amilcar Ferraz Farina 188.708.208-58

*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

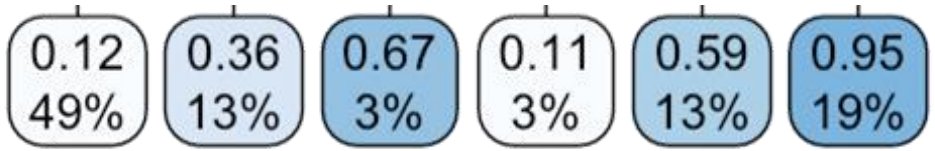
Árvore como diagnóstico

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

Especificidade: $\frac{TN}{TN+FP} = \frac{484}{484+65} = 72\%$

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

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



Anilcar Ferraz Farina 188.708.208-58

*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

Diagnóstico e pontos de corte

Corte	TP	FP	TN	FN
0,0 % |- 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%

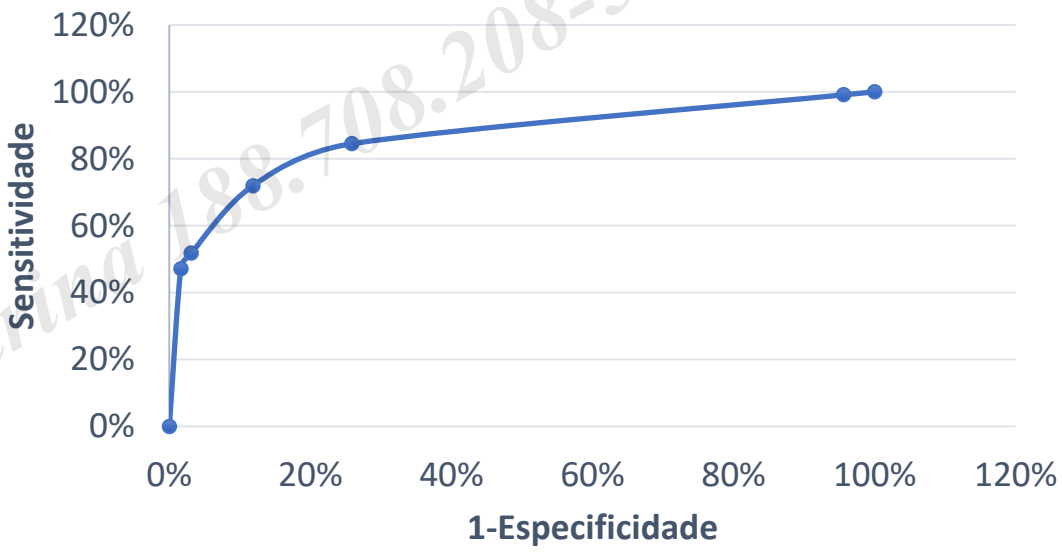
Amilcar Ferraz Farina 188.708.208-58

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

*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

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%



Amilcar Ferraz Fanna 188.708.208-58

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.

*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

Amilcar Ferraz Farias 188.108.208-58

ARNEM1 _script02-Algoritmo_avaliacao_overfitting

*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

Árvores de decisão:

Onde vivem?

O que são?

O que comem?

Predadores naturais

Marcos Roberto Farina 188.708.208-58

*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

O que é

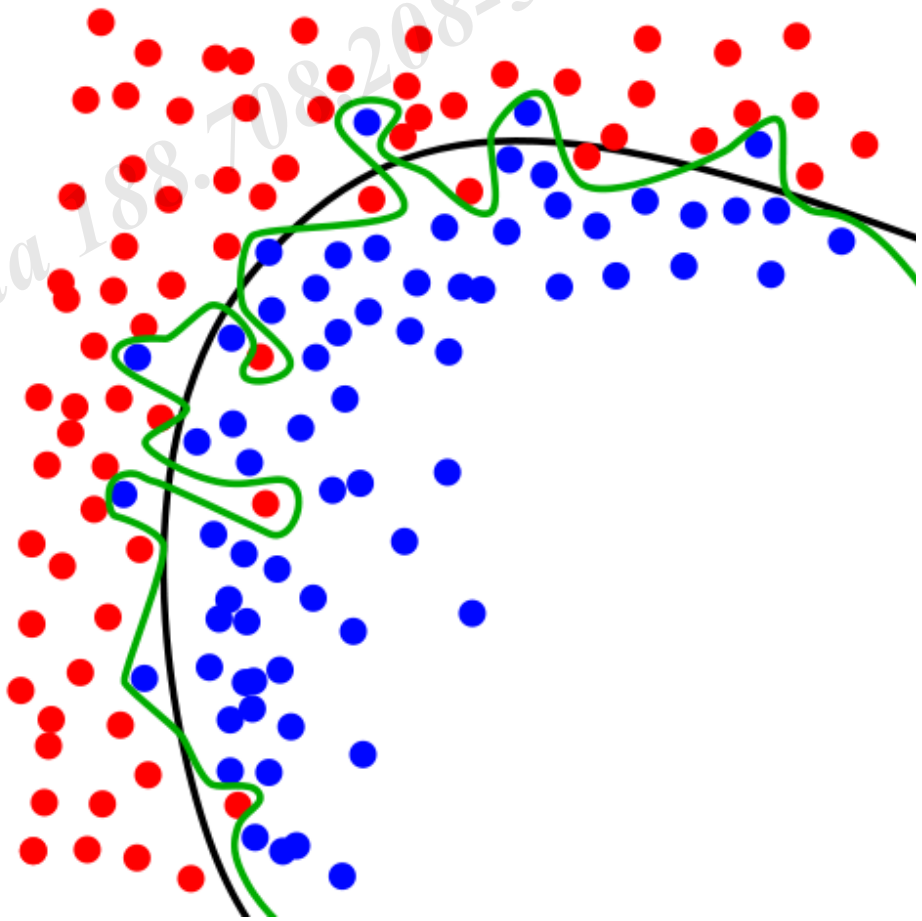
Como evitar

*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

Overfitting

- *Overfitting* é o fenômeno do modelo se ajustando a características muito particulares de amostra que não se repetirão em outras amostras.

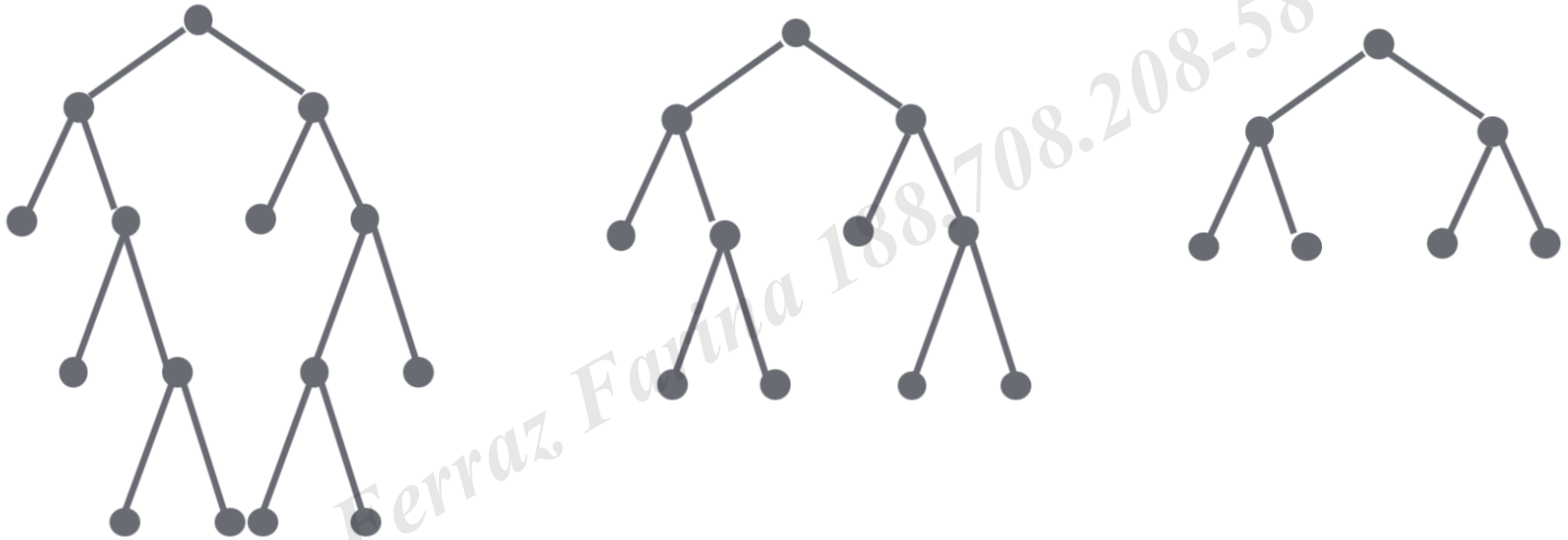
- Isso implica que a qualidade do modelo poderá ser bem menor na aplicação do modelo.



Amilcar Ferraz Farina 188.708.208-58

*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

Poda da árvore *(Prunning)*



Anilcar Ferraz Fanna 188.708.208-58

Acurácia		
Base de treino: 95%	Base de treino: 70%	Base de treino: 65%
Base de validação: 40%	Base de validação: 60%	Base de validação: 64%

Amostra de treino | Amostra de validação

*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

Estratégias de cross validation

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

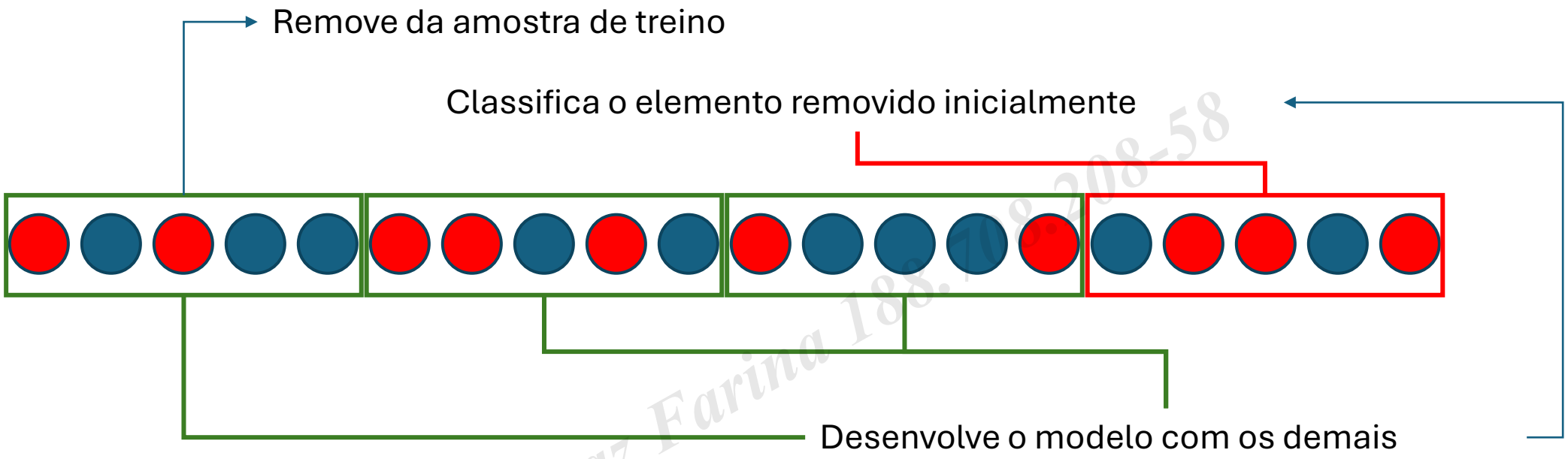
Há diversas técnicas de validação cruzada para se evitar esse efeito. No momento vou mencionar uma técnica clássica: dividir a base em Treino, Validação e Teste

Amilcar Heiraz Farina 188.708.208-58

Amostra de treino	Amostra de validação	Amostra de teste

*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

K-fold



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

Arnaldo Ferraz Farina 188.708.208-58

*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

K-fold

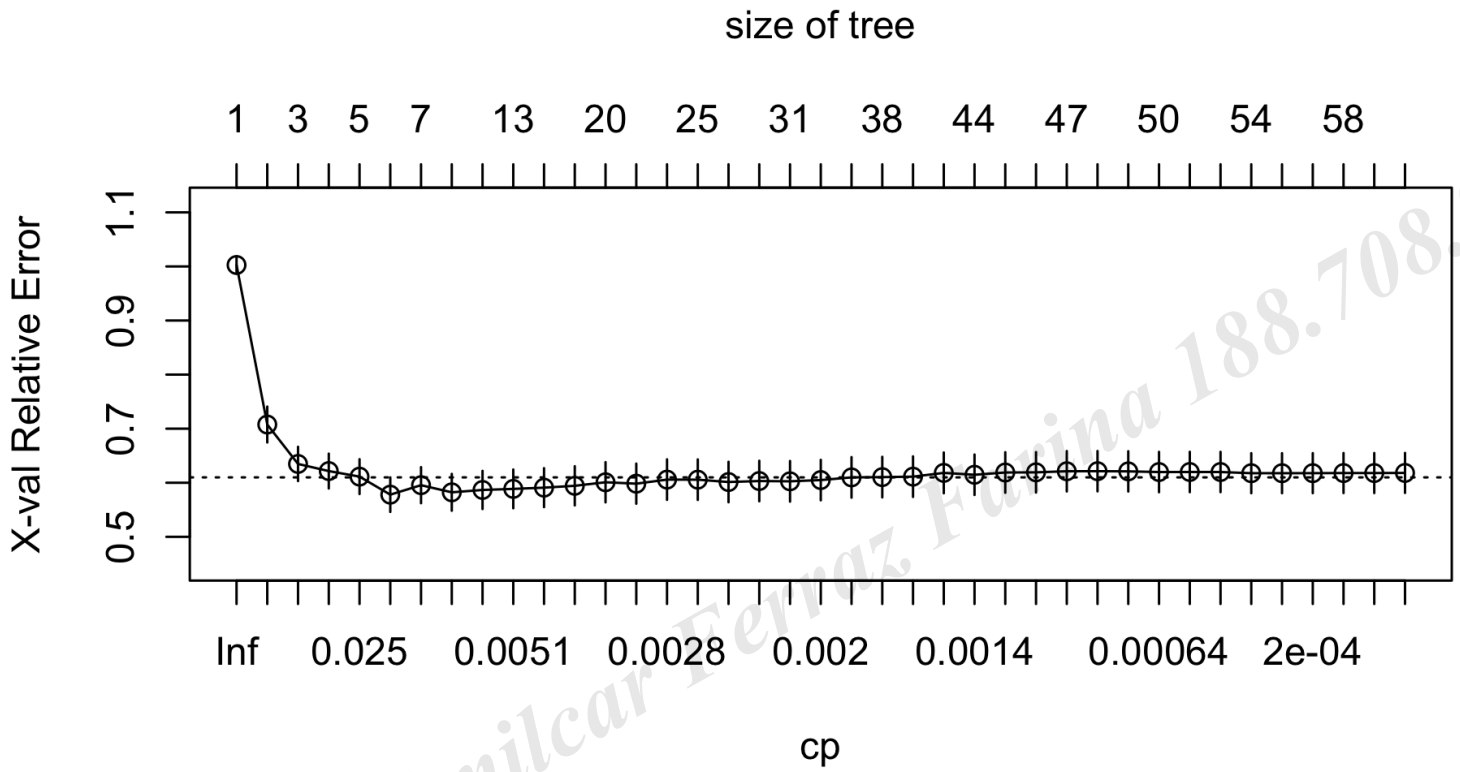
Tipicamente, fazemos o mesmo para variações do modelo para otimizar hiperparâmetros.

	Acurácia 1	Acurácia 2	Acurácia 3	Acurácia 4	**Acurácia Média**
Modelo 1	62%	58%	61%	59%	**60%**
Modelo 2	50%	51%	49%	47%	**49%**
Modelo 3	72%	68%	71%	75%	**72%**

Amilcar Ferraz Farina 188.708.208-58

*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

Post-prunning com crossvalidation



printcp(arvore)

Amilcar Ferraz Farina 188.708.208-58

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*.

*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

Amilcar Ferraz Farias 188.108.208-58

ARNEM1 _script02-Algoritmo_avaliacao_overfitting

*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

Exercício para casa

script03

Amilcar Ferrari Sarina 188.708.208-58

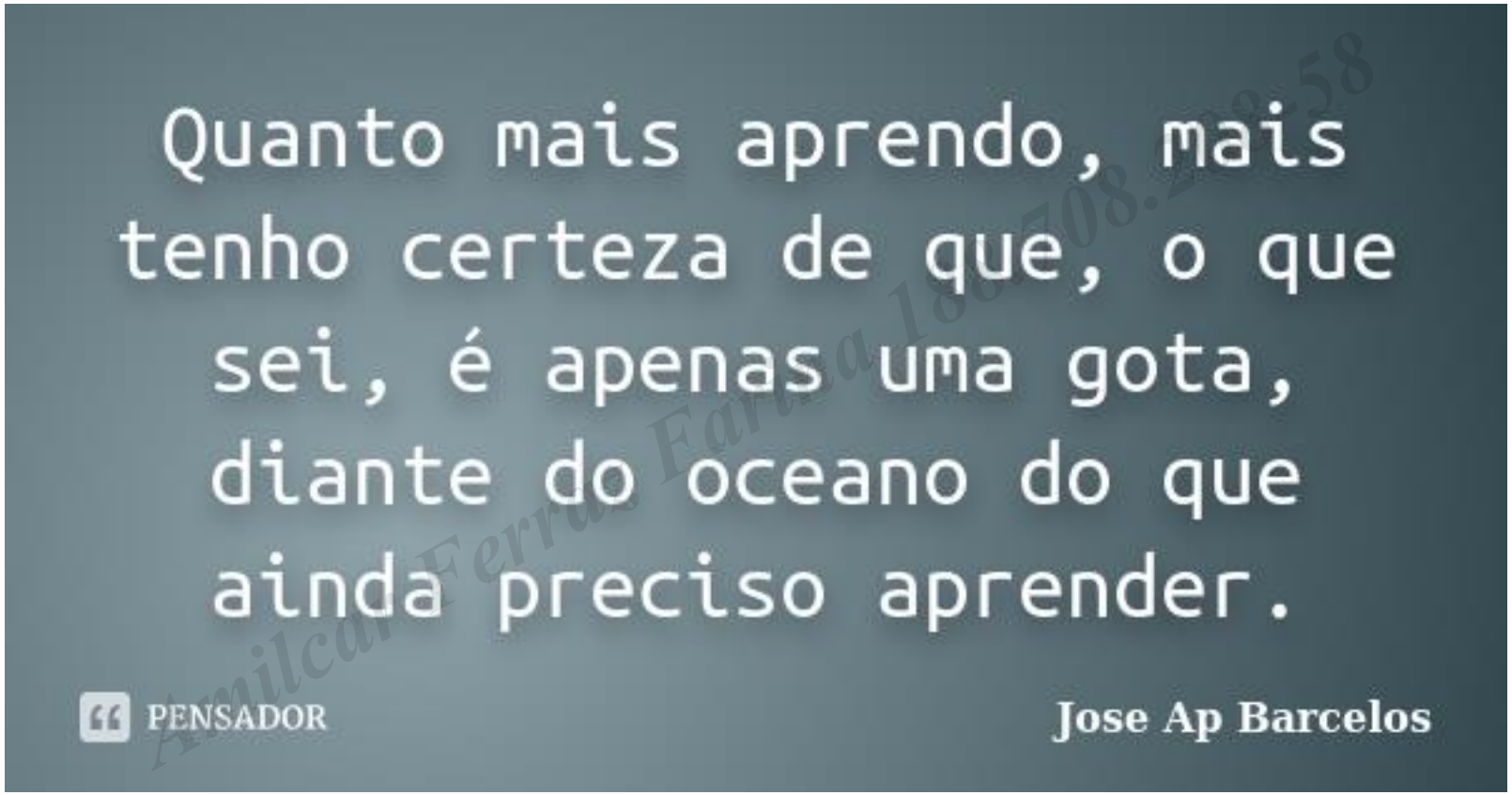
*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.
Proibida a reprodução, total ou parcial, sem autorização. Lei nº 9610/98

Conclusão

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

Amilcar Ferraz Farina 108.708.208-58

*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



*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

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

Amilcar Ferraz Farina 188.708.208-58

*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

Marina 188.708.208-58

Por hoje é só!!!

*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

MBA USP ESALQ

Obrigado!

Professor João Serrajordia | linkedin.com/in/joao-serrajordia

Amilcar Ferraz Farina 188.208.208-58