MBA USP ESALQ

Other Machine Learning Models I

João F. Serrajordia R. de Mello

Introduction

João Fernando Fernando Serrajordia Rocha de Mello – (Juka)

Professional journey

Credit modeling in large banks
Telecom
Development of models / Validation of models

Teaching in data science
Consultancy in data science
Executive *outsourcing*

ACADEMIC

IME
USP

Bachelor's Degree

Master in Statistcs

MBA USP ESALQ


Decision Trees

You will need...

Preparations

- Open R
- Import libraries
- Electronic sheet
- Something to take your notes

Decision Trees:

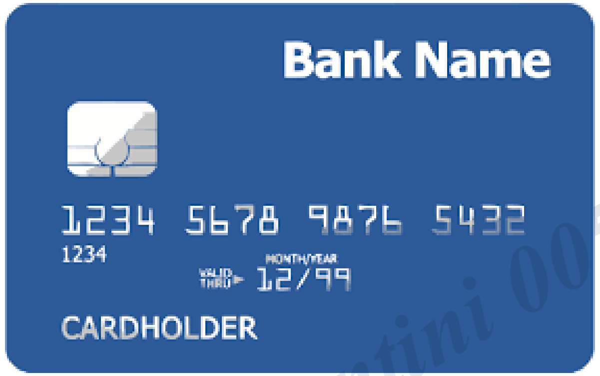
Where do they live? What are they?

What do they eat?

Natural Predators

Predictive and classification problems

What is the efficacy of a vaccine?



Will the customer pay the loan?

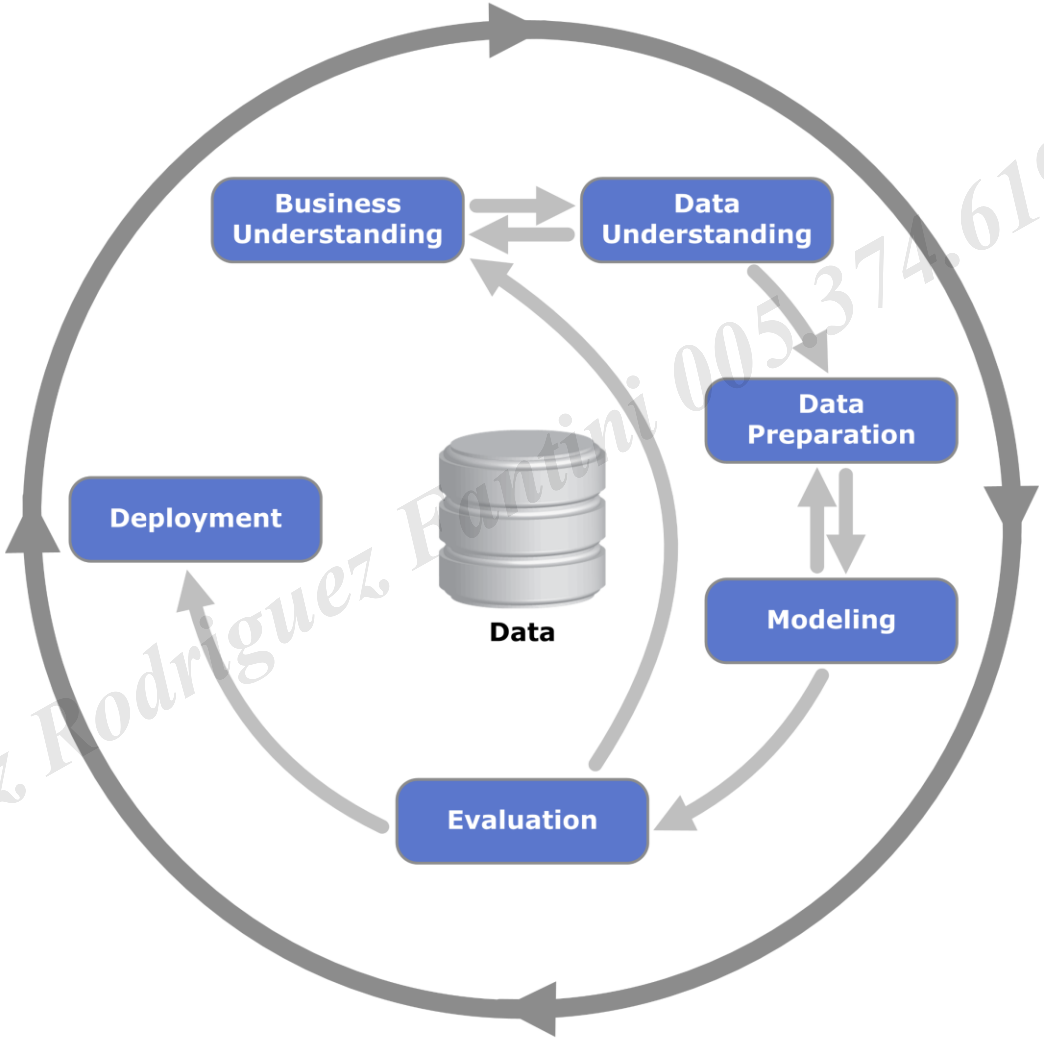
How much oil is in the well?

Will the customer buy my product?

What is the person doing?

How green is this vehicle?

CRISP-DM

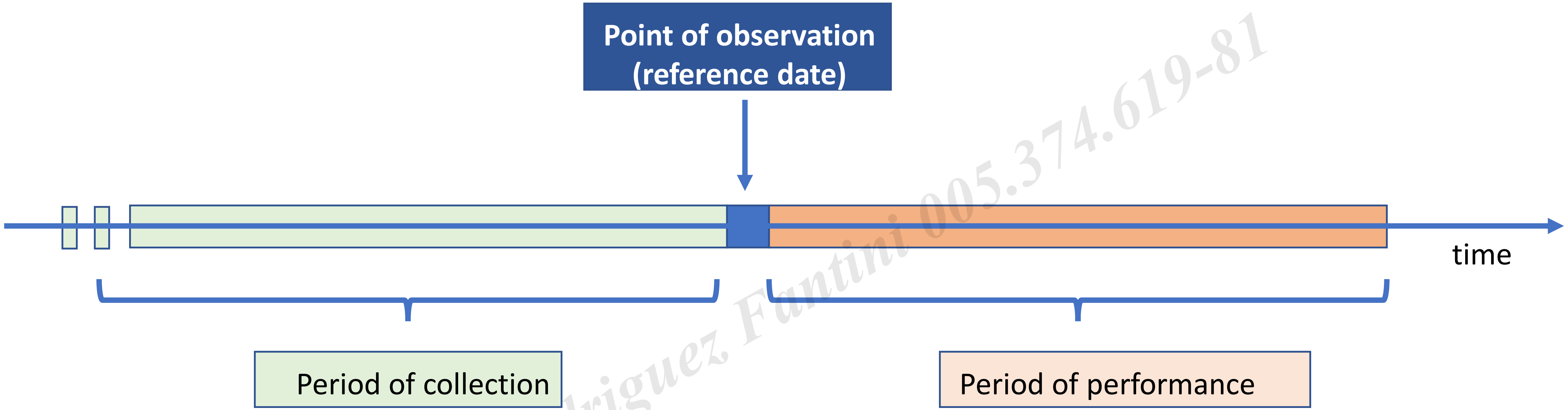


Source: https://www.the-modeling-agency.com/crisp-dm.pdf

Predictive models

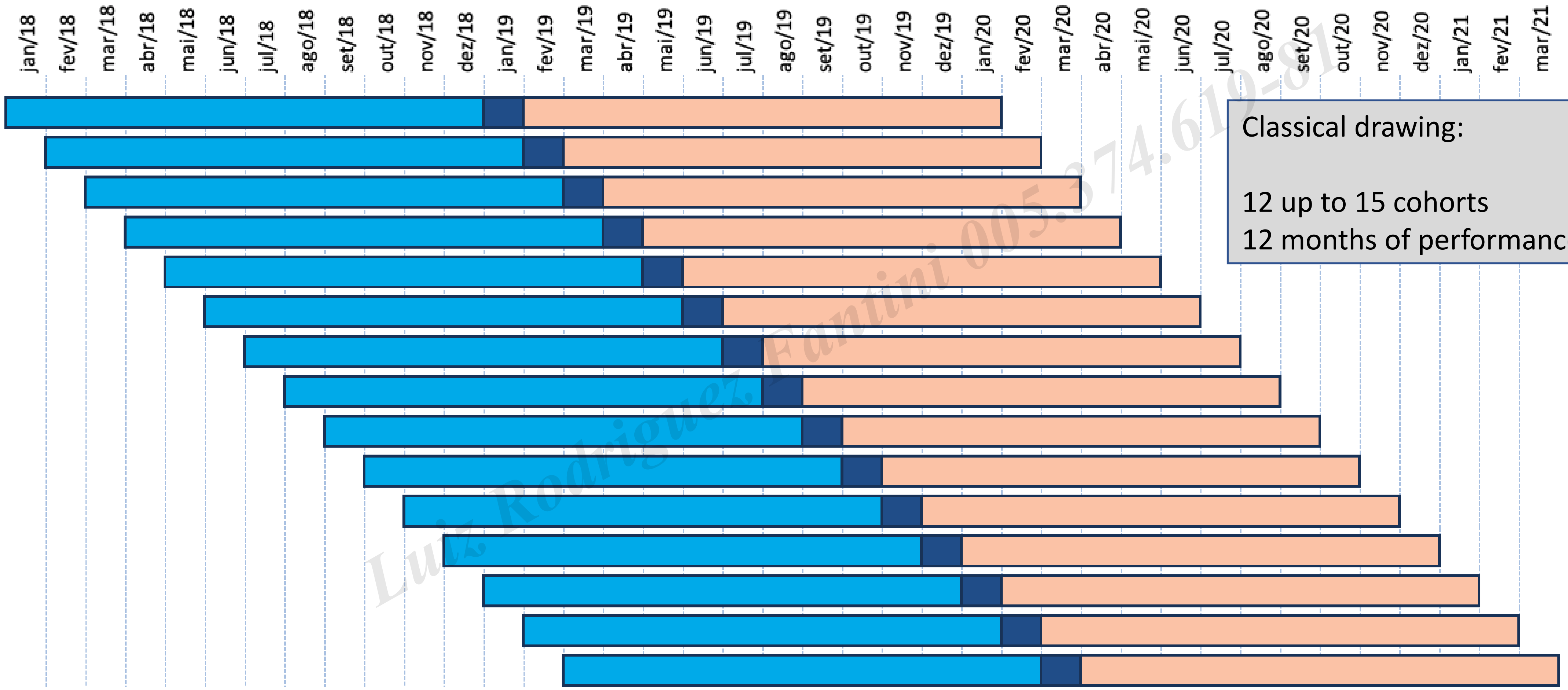
How is this?

Cohort



Example of drawing a sample for predictive model

Model Drawing



Classical drawing:

12 up to 15 cohorts
12 months of performance

Algorithms classification

Supervised

- Regression
- GLM
- GLMM
- Support vector machines
- Naive Bayes
- K-nearest neighbors
- Neural Networks
- Decision Trees

Unsupervised

- K-Means
- Hierarchical methods
- Gaussian Mixture
- DBScan
- Mini-Batch-K-Means

We are here!

Algorithms classification

Continuous response

- Regression
- GLM
- GLMM
- Support vector machines
- K-nearest neighbors
- Neural Networks
- Regression Trees

Discrete response

- Logistic Regression
- Classification trees
- Neural Networks
- GLM
- GLMM

We are here!

Algorithms classification

Machine Learning
Methods

- Regression
- GLM
- GLMM
- ANOVA

Machine Learning
Statistics Methods

- Decision Trees
- Bagging
- Boosting
- K-NN
- Neural Networks
- Support Vector Machines

We are here!

Our problem: classify survivors

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

Reflections on the database

Population

- ~ 2,200 people
- ~ 1,300 passengers
- More than 1,500 deaths

Sample

- 891 people
- 549 non-survivors
- 342 survivors

Objectives of algorithm

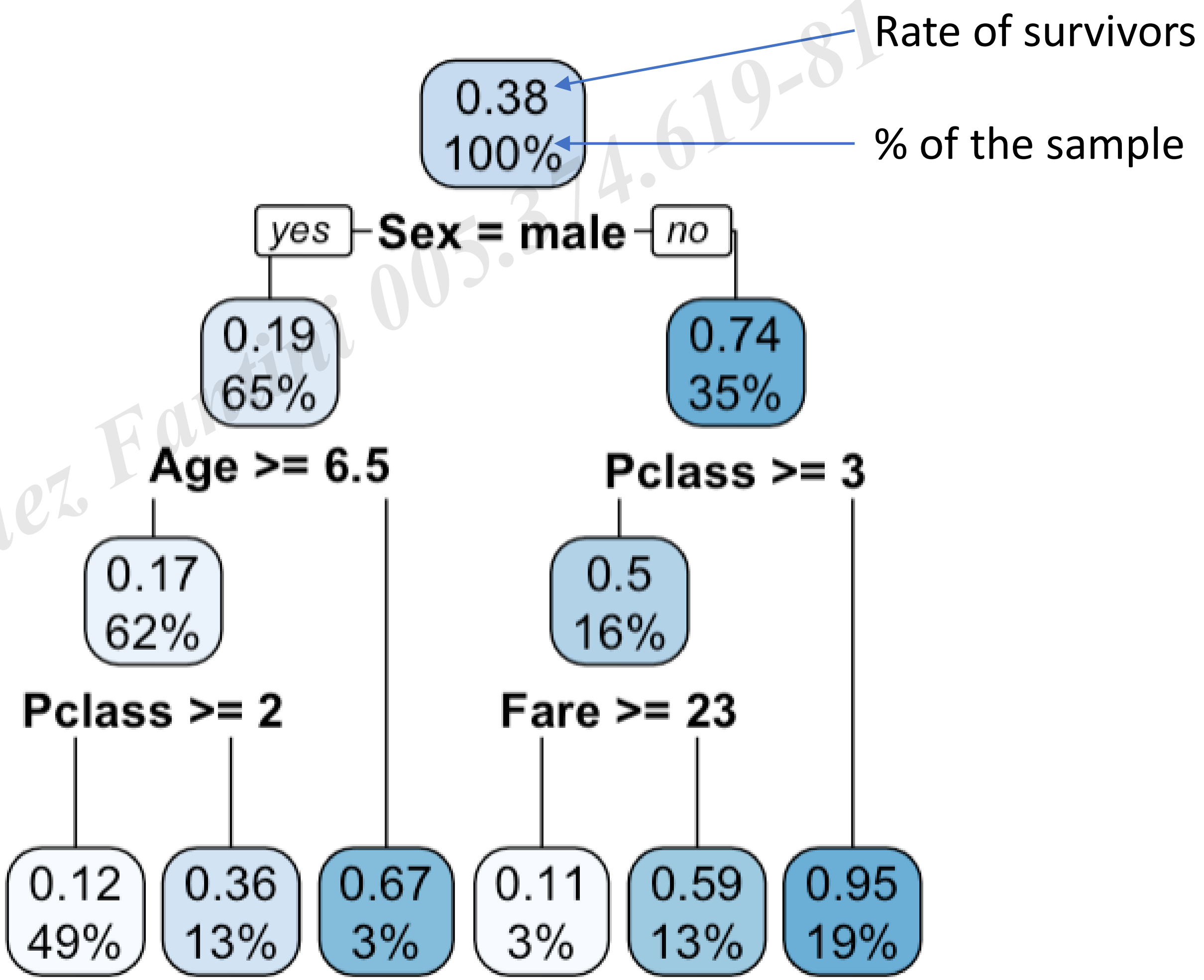
- To classify the response variable as well as possible
  - ... Through segmentations
  - ... Using explanatory variables
- To obtain insights
  - ... From relations between the response variable and the explanatory variable
  - ... To explore interactions

OMML1_script01-Primeiro_contato_com_arvores.R

What is a decision tree?

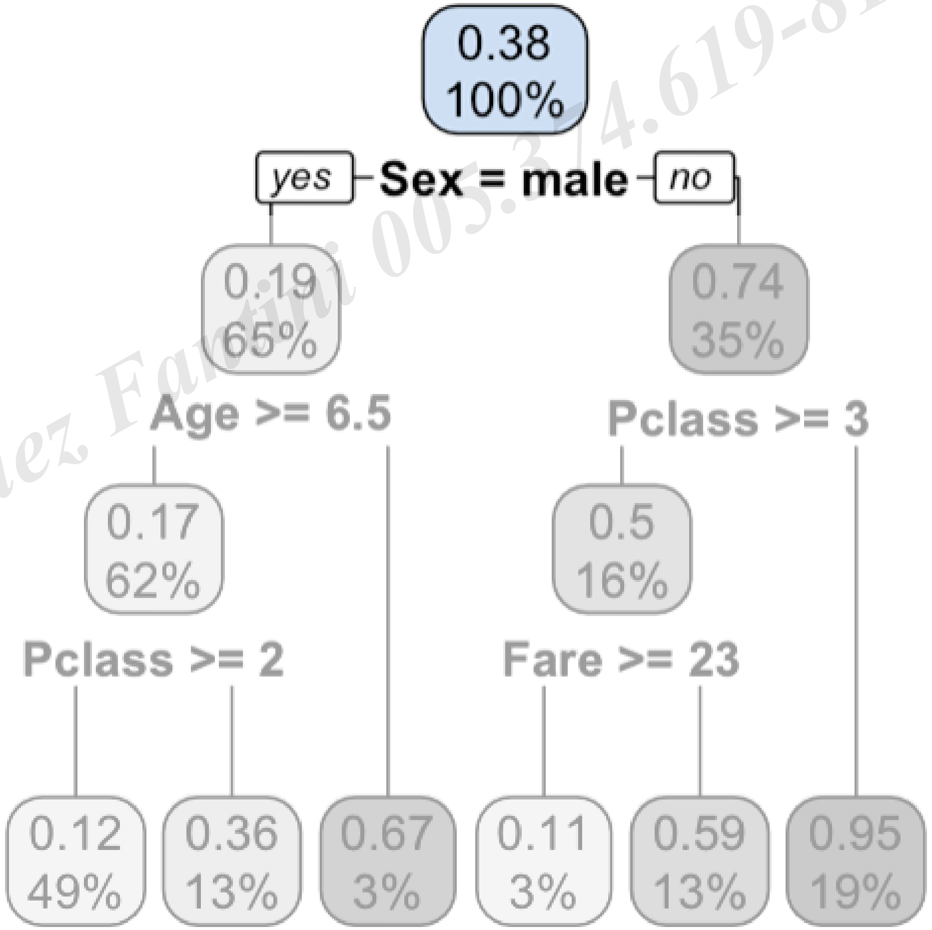
Decision tree is:

A sequence of binary segmentations
That aims the homogeneity of the response variable



What is a decision tree?

Initially we have 891 passengers of which
342 survived (38%)
549 did not survive

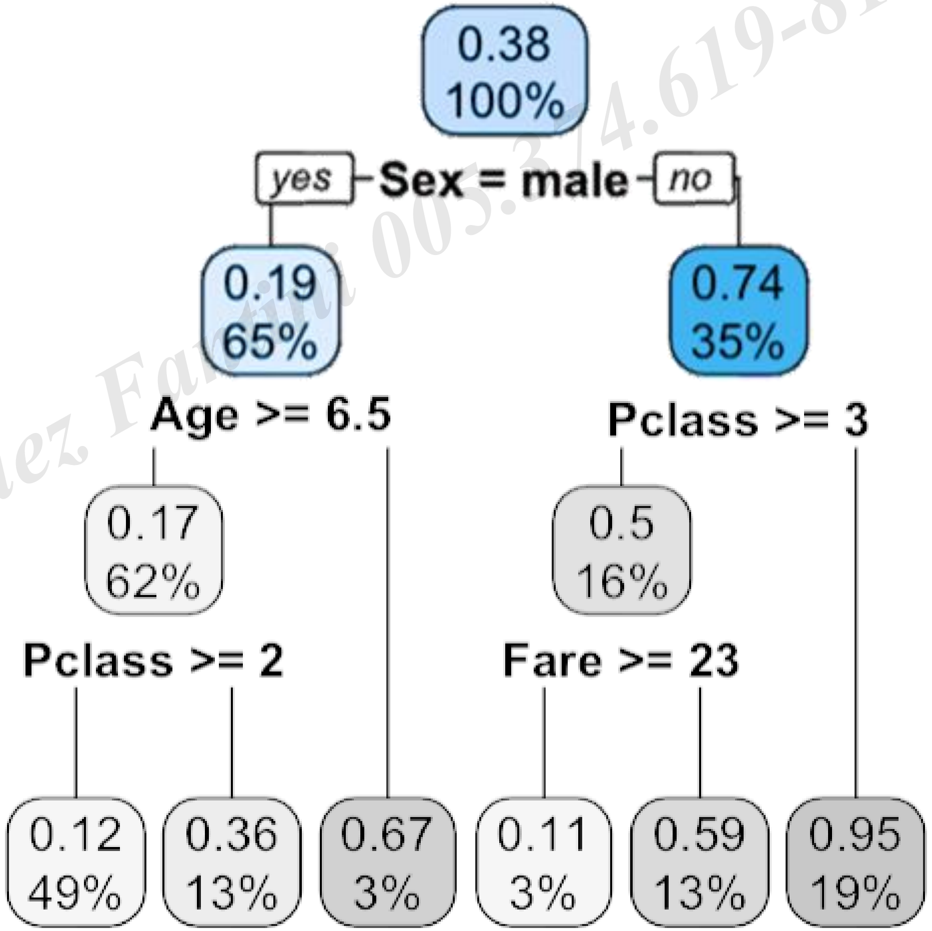


What is a decision tree?

We can segment from the 891:

577 men (65%) of which
109 survived (19%)
468 did not survive

314 women (35%) of which
233 survived (74%)
81 did not survive



MBA USP ESALQ

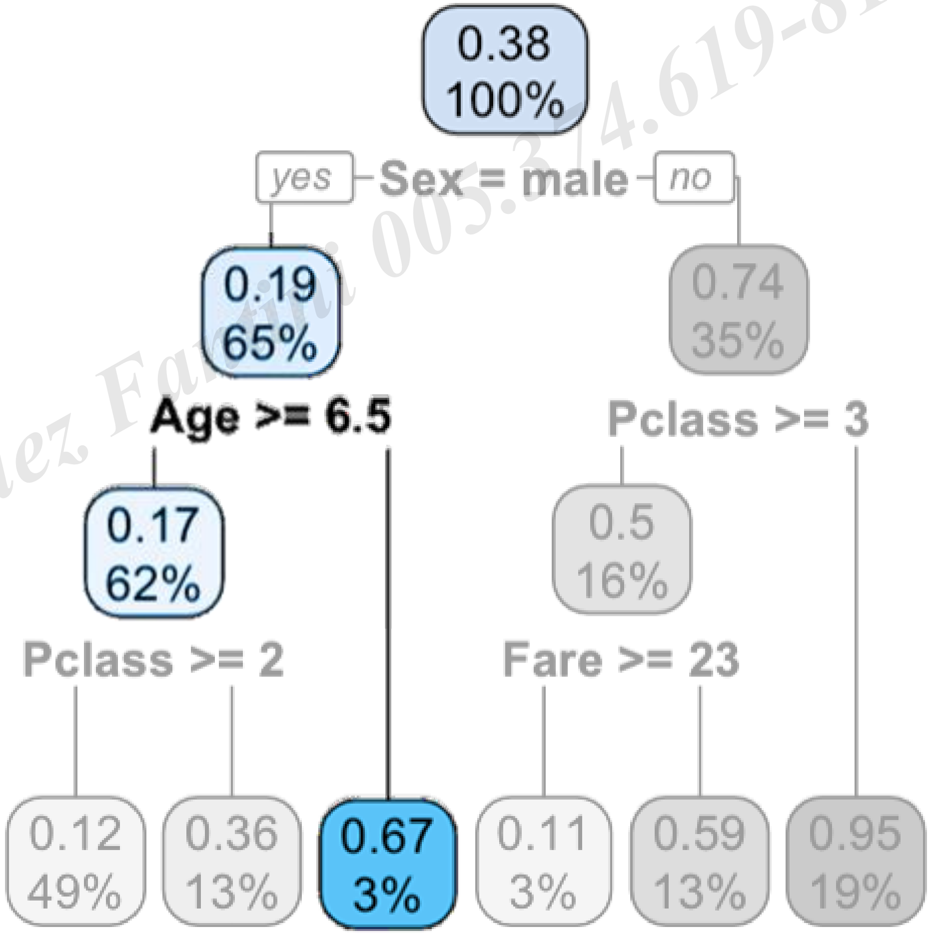
What is a decision tree?

We can segment from the 891:

577 men that we segment in:

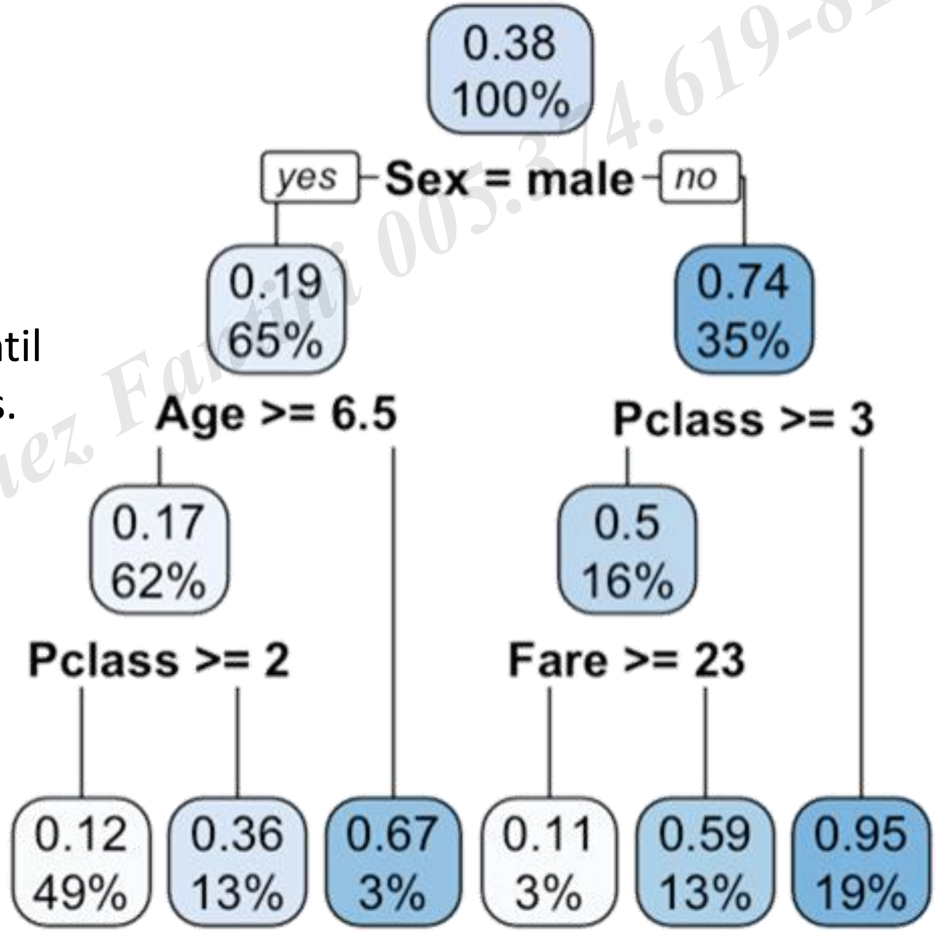
24 kids (< 6, 5 years) of which
16 survived (67%)
8 did not survive

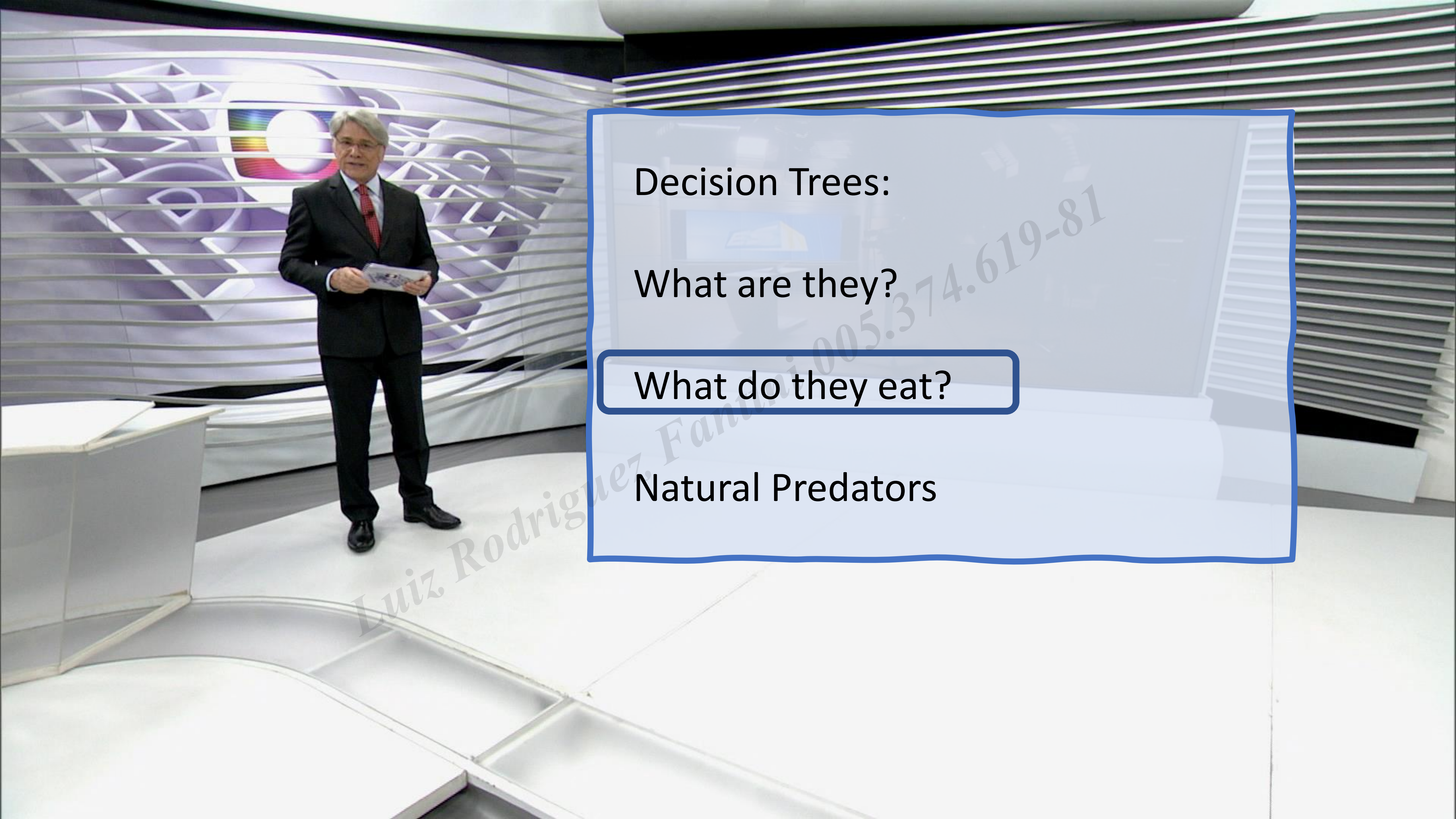
533 adults (>=6, 5 years) of which
93 survived (17%)
553 did not survive



What is a decision tree?

And, then, we continue to "split" the sample until "it is not worth it anymore" to make more splits.




Decision Trees:
What are they?
What do they eat?
Natural Predators

Definitions of impurity

- Gini
- Shannon Entropy

How does the tree select the best split?
Using a metrics of "impurity"

Gini's index

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

- Maximum impurity with uniform distribution
- Minimum impurity in the total concentration

Entropy

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

Information gain:

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

- Maximum impurity with uniform distribution
- Minimum impurity in the total concentration

Basic Algorithm

1. Seek the best binary rule for each variable.
2. Seek to apply the best segmentation among all variables
3. Recursively, for each sheet, repeat the steps 1 and 2 until a stopping rule is reached.

Interactive web implementation:

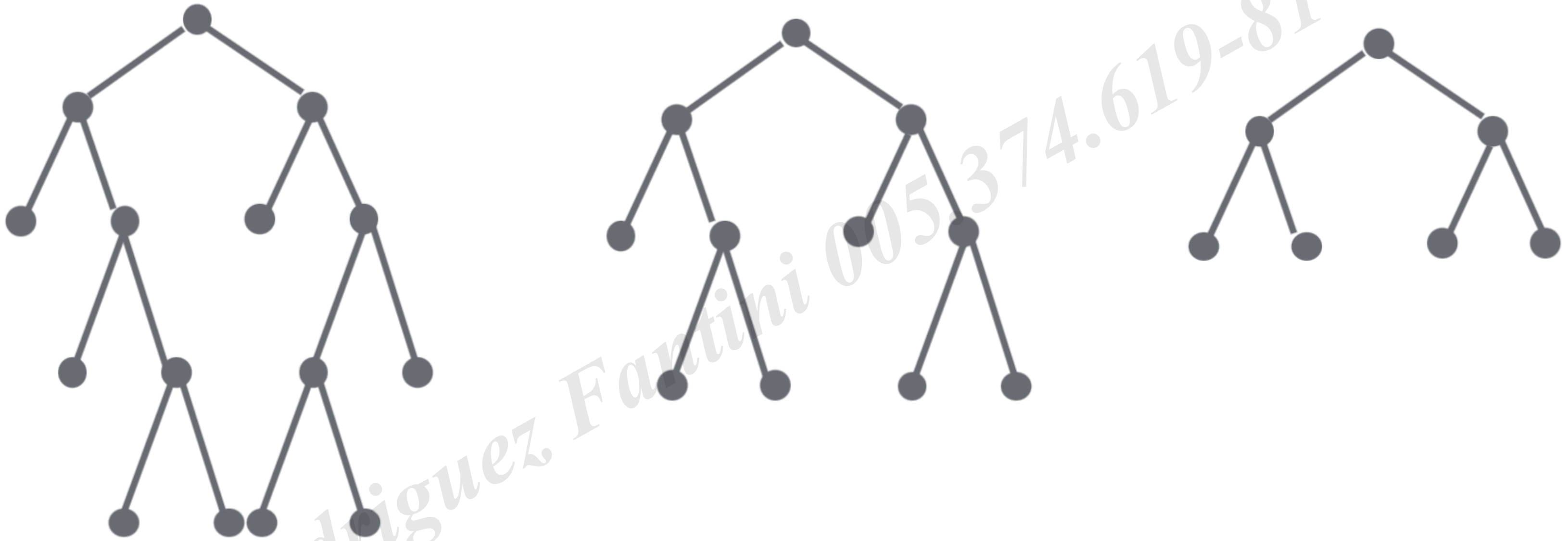
https://rawgit.com/longhowlam/titanicTree/master/tree.html

Hyperparameters

They are parameters that control the algorithm as:

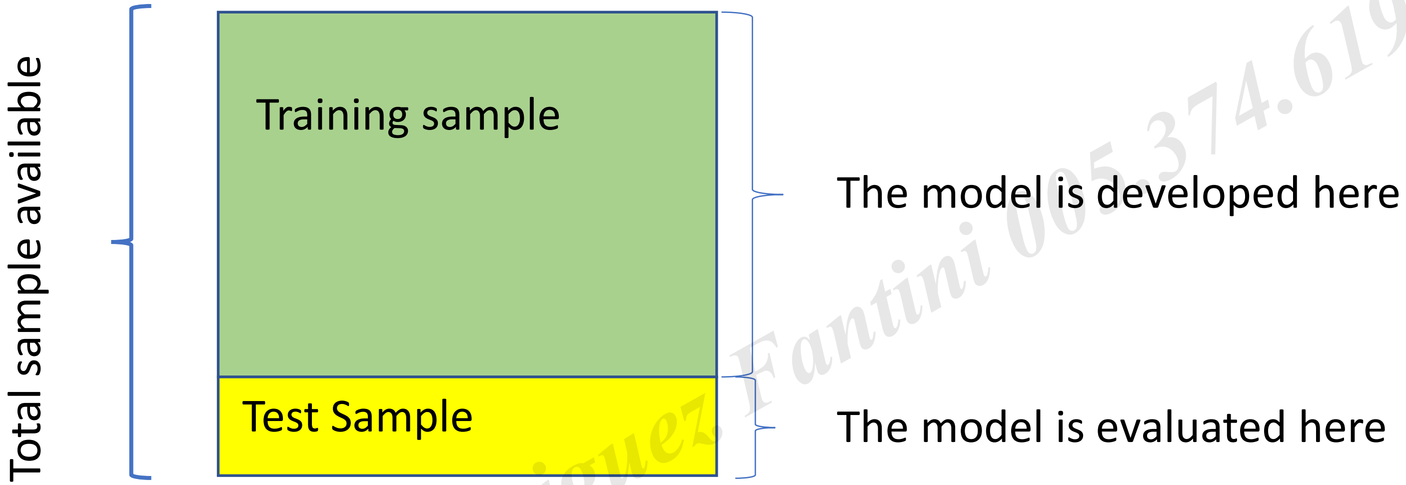
1. Minimum number of observations per sheet
2. Maximum depth
3. Cost of complexity

Cost of complexity



Cost of complexity	Low	Moderate	High

Cross validation



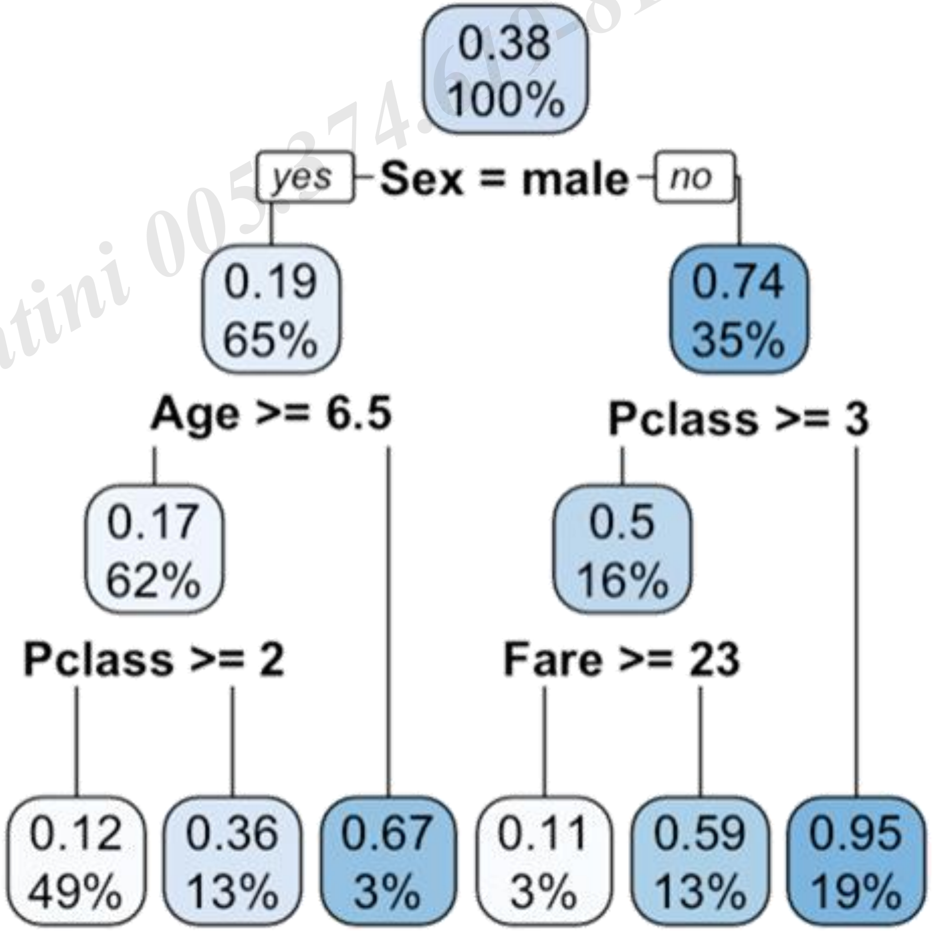
The most simple strategy is to split the basis into training and test. We develop the model in the training base, and evaluate the test base.

OMML1 _script02-Algoritmo_avaliacao_overfitting

The tree as a classifier

Requisites:

To have all variables.



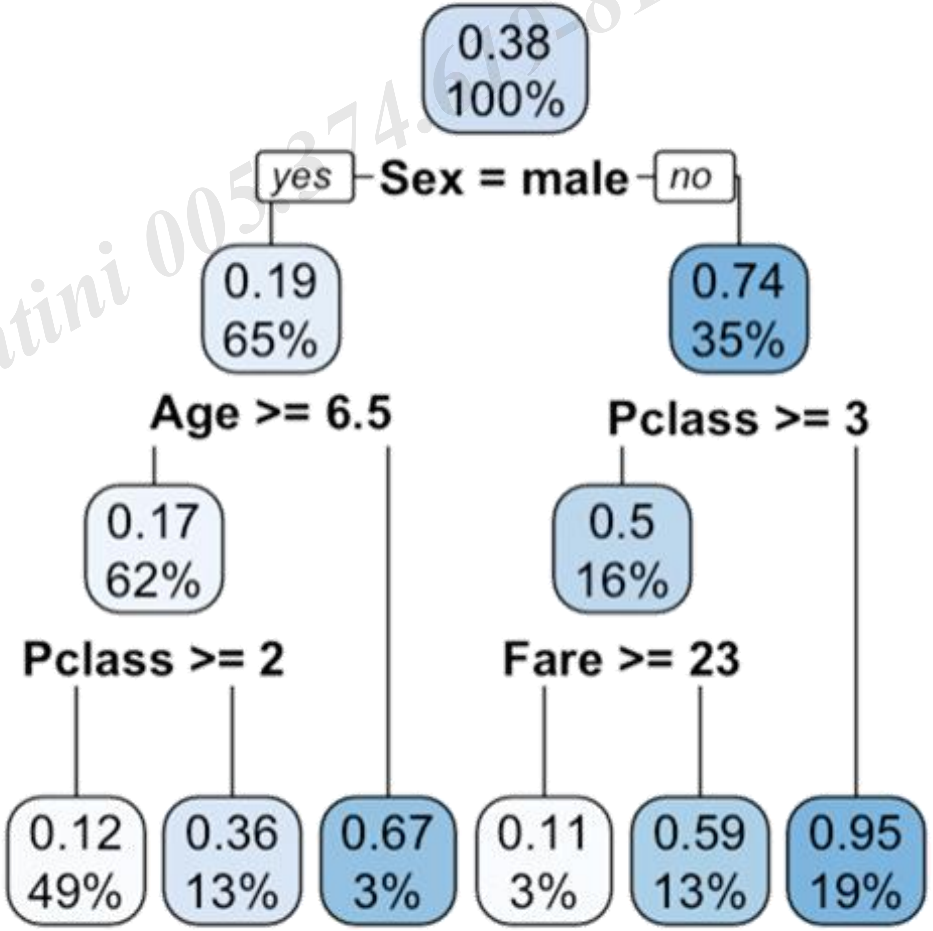
The tree as a classifier

Probability of the F. event:

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

P(S|F) - probability of success of the F sheet

N_f - it is the number of individuals on F sheet
N_f^S - it is the number of survivors on F sheet



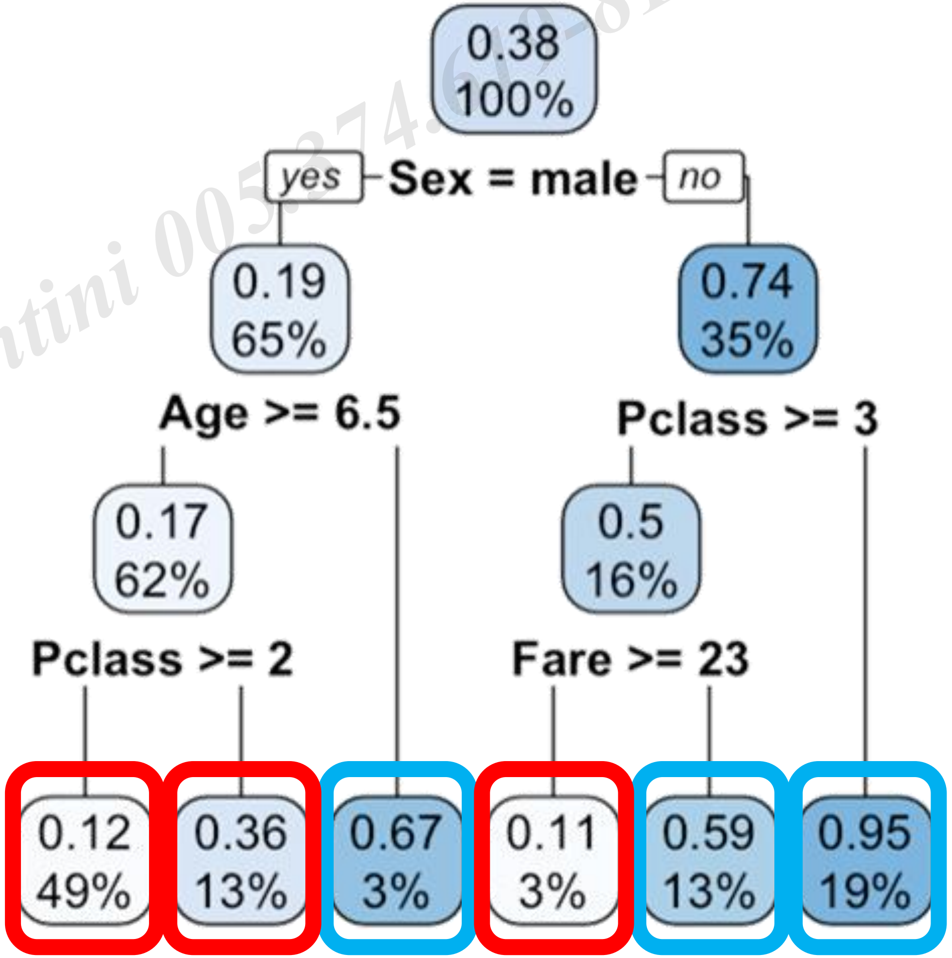
The tree as a classifier

Classification:

Standard classification:

Survivor: $P(S|F) \geq 50\% \implies C(F) = \text{“}Y\text{”}$
No survivors:$P(S|F) < 50\% \implies C(F) = \text{“}N\text{”}$

Predicted value	True Value	
	0	1
0	484	96
1	65	246



Evaluation of the model

- Accuracy:

Hits on attempts

Predicted value	True Value	
	0	1
0	484	96
1	65	246

In this example:

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

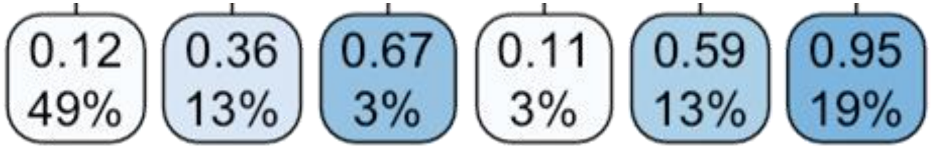
Trees as diagnosis

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

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

Predicted value	True Value	
	0	1
0	484	96
1	65	246

Predicted value	True Value	
	0	1
0	TN	FN
1	FP	TP



Diagnosis and cutoff points

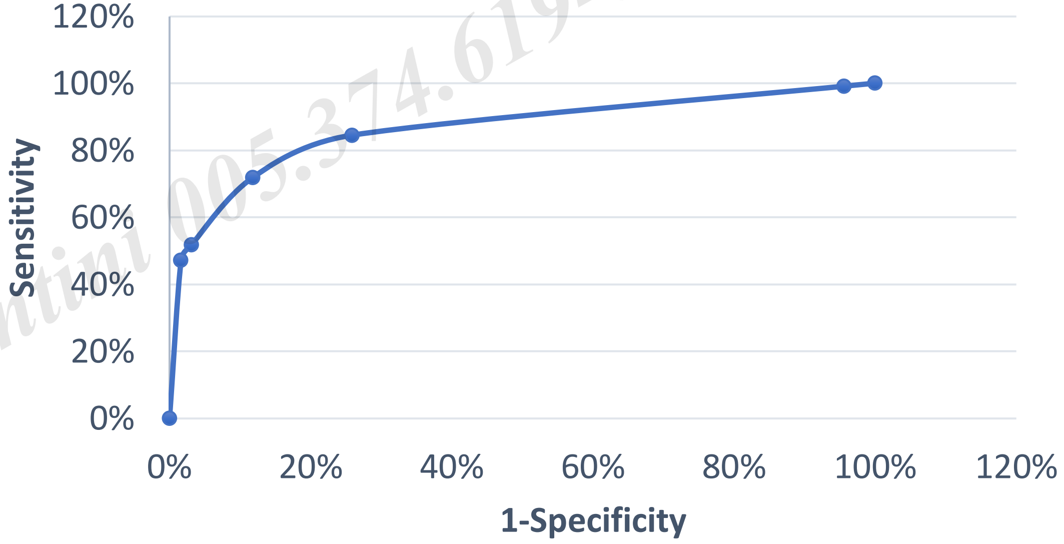
CUT	TP	FP	TN	FN
0% - 11.1%	342	549	0	0
11.1% - 11.5%	339	525	24	3
11.5% - 35.8%	289	142	407	53
35.8% - 58.9%	246	65	484	96
58.9% - 66.7%	177	17	532	165
66.7% - 94.7%	161	9	540	181
94.7% - 100%	0	0	549	342

Accuracy	Specificity	1-Specificity	Sensibility
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%

For each cutoff point, we have a confusion matrix.
In this case, we have 8 possible matrices with the trained tree.

ROC Curve

CUT	1-Specificity	Sensibility
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%



ROC curve is a graphical plot of 1-Specificity on the X-axis by Sensitivity on the Y-axis, obtained for each possible cutoff point of the classifier.

OMML1 _script02-Algoritmo_avaliacao_overfitting

Decision Trees:

What are they?

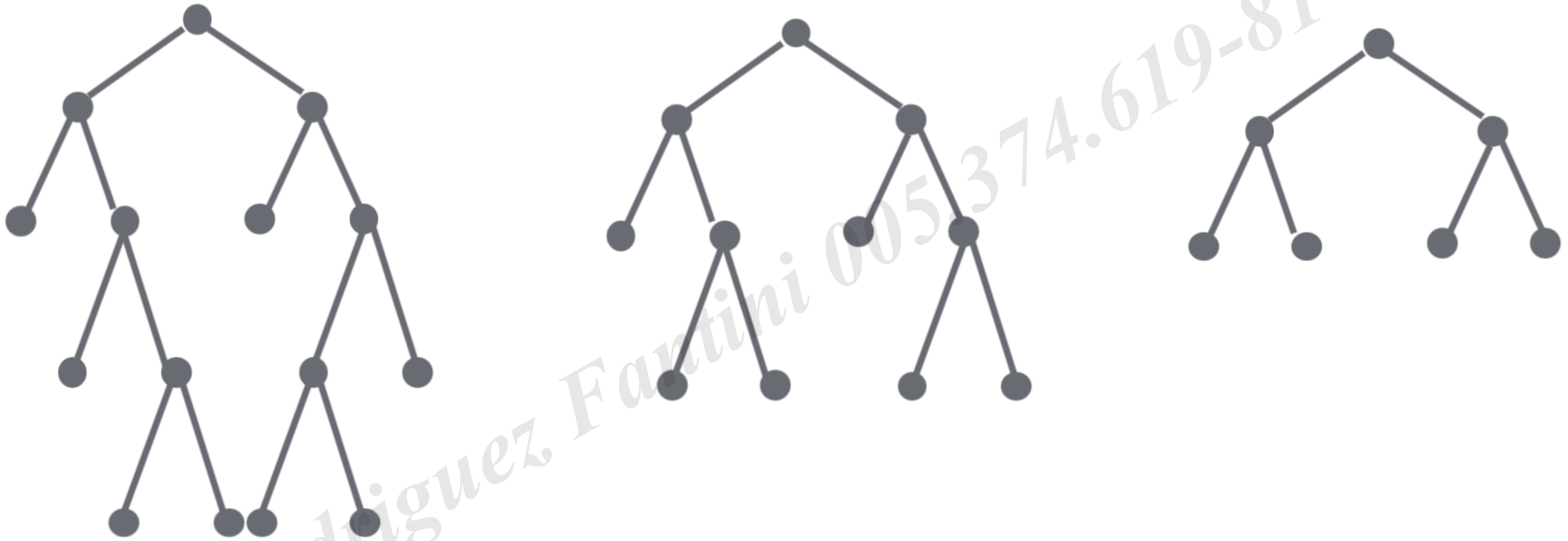
What do they eat?

Natural Predators

What is it?

How to avoid it?

Pruning



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

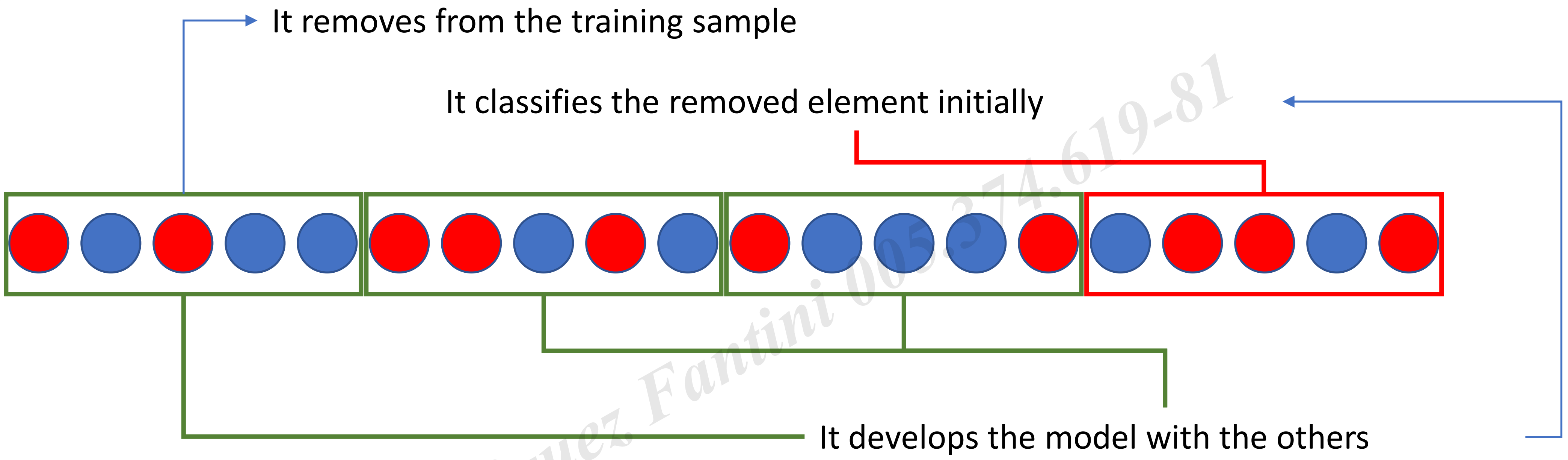
Cross validation strategies

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

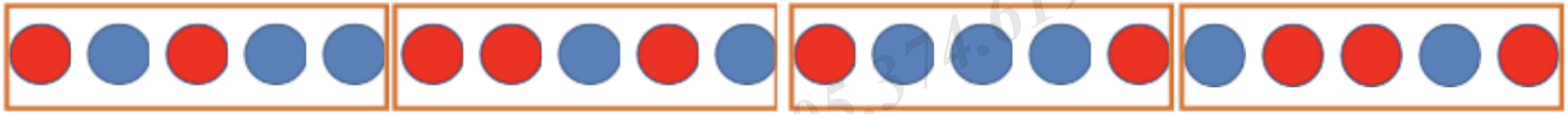
K-fold



- We divide the base into sub-samples k
- For each sub-sample:
    - We remove the sub-sample as validation
    - We train the model with the remaining observations
    - We use this model to classify the removed sub-sample
    - We evaluate the metrics of the model's performance
- We calculate the average of the metrics of the model's performance

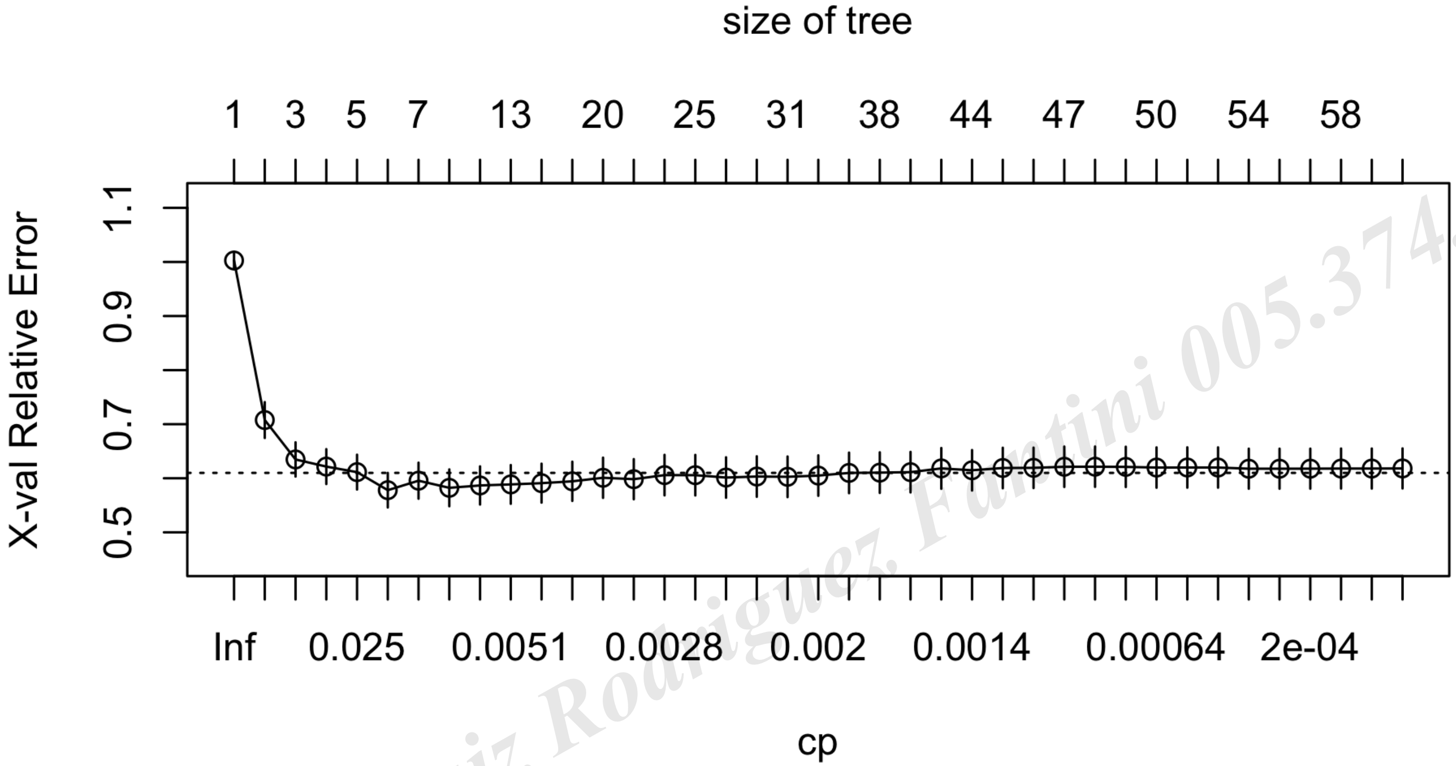
K-fold

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%**

Post-pruning with cross validation



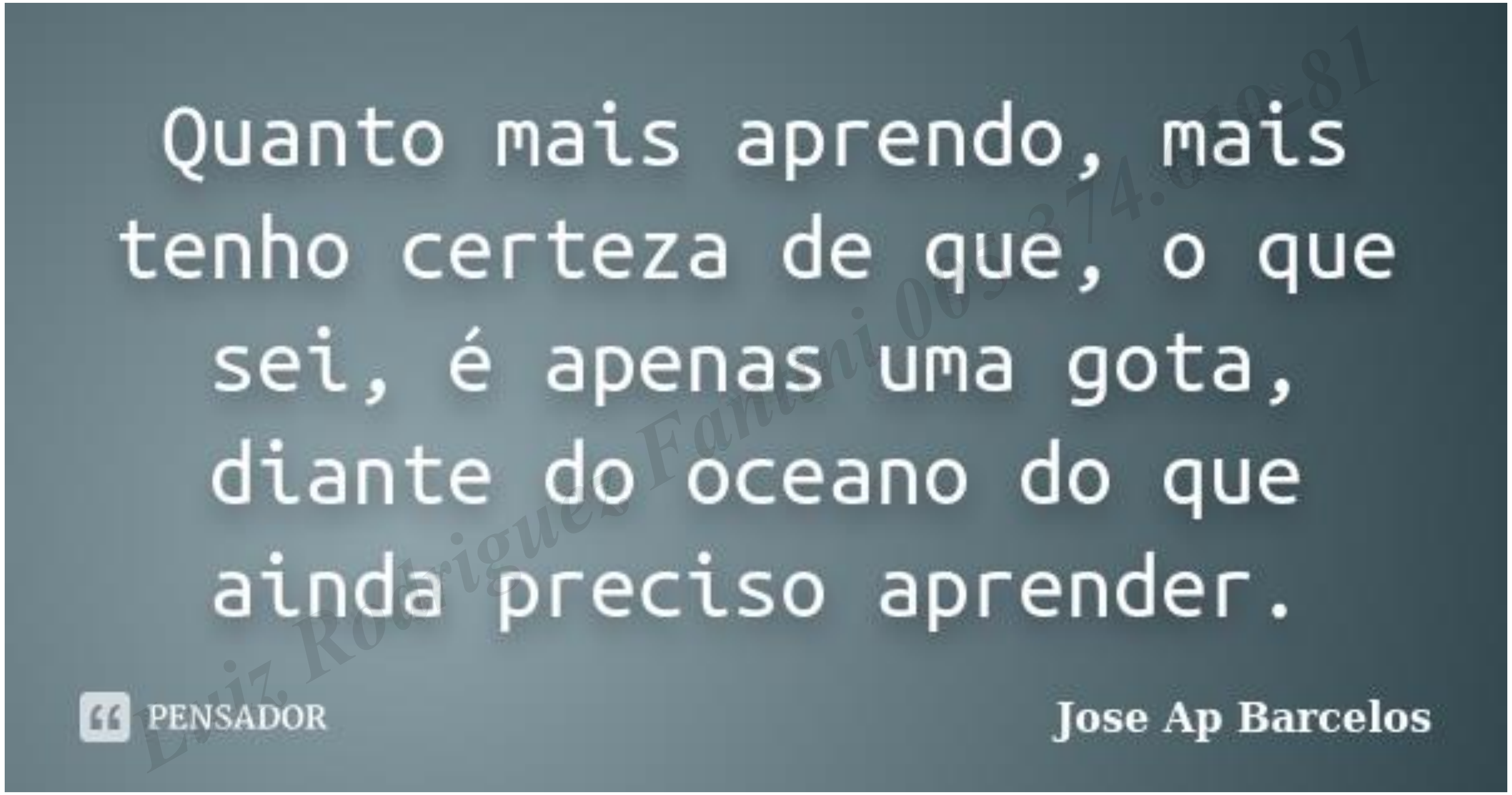
printcp(arvore)

R performs the pruning of the tree by performing a k-fold to optimize the CP (complexity path), a parameter that summarizes the complexity of the tree. This is made with a *k-fold*.

OMML1 _script02-Algoritmo_avaliacao_overfitting

Conclusion

- Robust, interpretable, flexible
- Without probabilistic assumptions
- It's necessary *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

That's it for today ;)

linkedin.com/in/joao-serrajordia

Famous Algorithms

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

Interesting stack overflow on this:

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