

MBA
USP
ESALQ

*Otros modelos de Machine
Learning I*

João F. Serrajordia R. de Mello

Presentación

João Fernando Serrajordia Rocha de Mello – (Juka)

Trayectoria profesional

Modelado de crédito en grandes bancos

Telecom

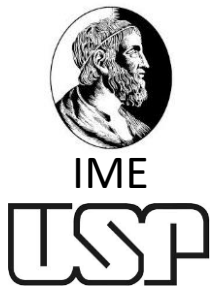
Desarrollo de modelos / Validación de modelos

Docencia en ciencia de datos

Consultoria en ciencia de datos

Outsourcing ejecutivo

Académico



BACHILLER EN ESTADÍSTICA

MASTER EN ESTADÍSTICA





Árboles de decisión

Va a necesitar de...



Preparativos

- Abrir R
- Importar las bibliotecas
- Planilla electrónica
- Algo para hacer sus anotaciones



Árboles de decisión:

¿Dónde viven? ¿Qué son?

¿Qué comen?

Predadores naturales

Problemas de predictivos y de clasificación



¿Cuál es la eficacia de una vacuna?



¿El cliente pagará el préstamo?



¿Cuánto petróleo tiene el pozo?



¿El cliente va a comprar mi producto?

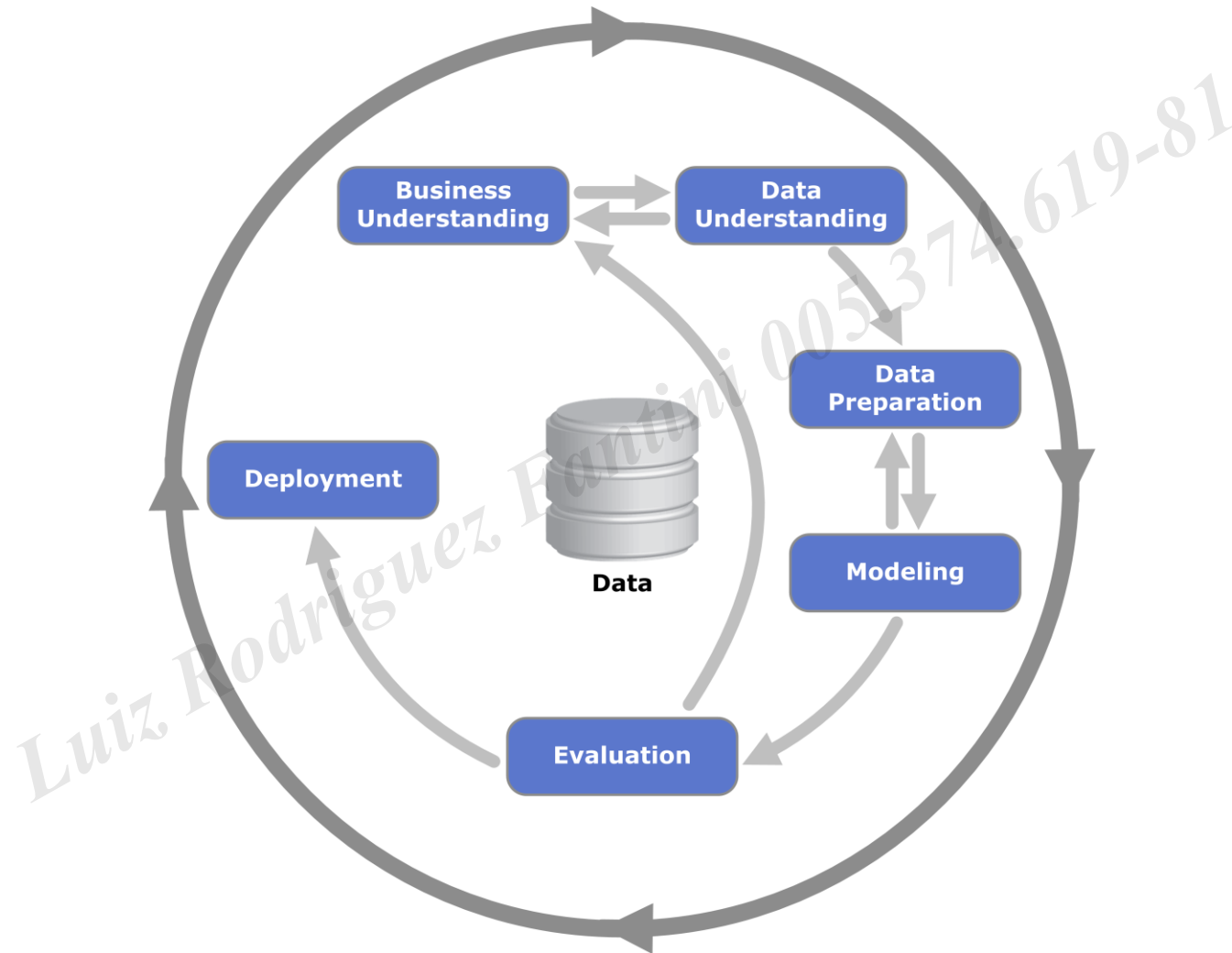


¿Qué está haciendo la persona?



¿Cuán ecológico es ese vehículo?

CRISP-DM



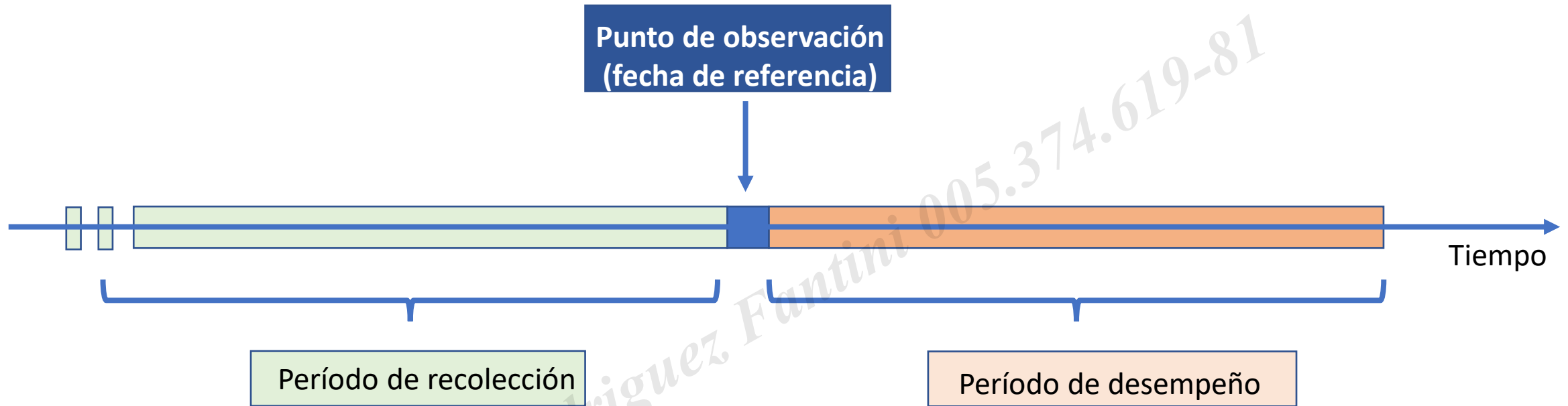
Fuente: <https://www.the-modeling-agency.com/crisp-dm.pdf>



Modelos predictivos

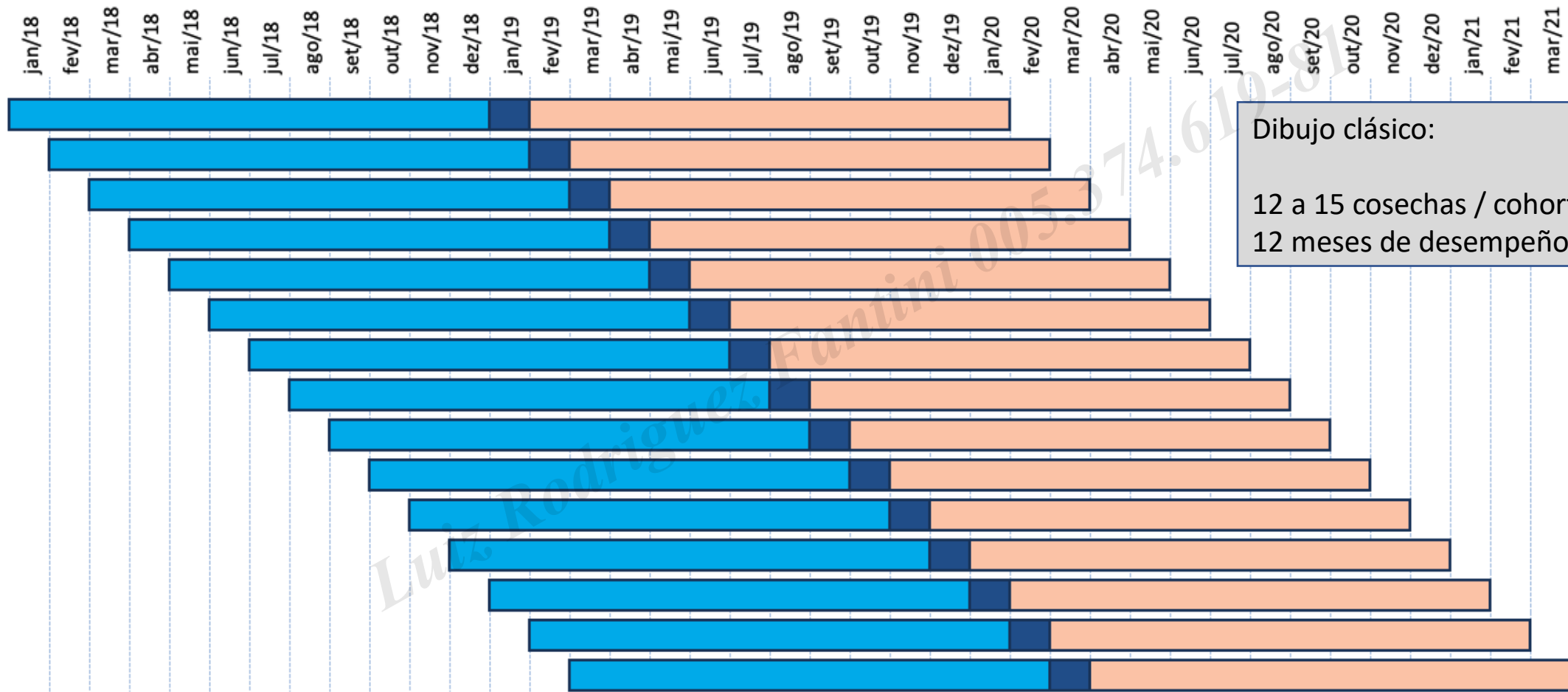
¿Cómo es eso?

Dibujo de cosecha (o cohorte)



Ejemplo de dibujo de muestreo para modelo predictivo

Dibujo del modelo



Dibujo clásico:

12 a 15 cosechas / cohortes
12 meses de desempeño

Clasificación de los algoritmos

Supervisados

- Regresión
- GLM
- GLMM
- Support vector machines
- Naive Bayes
- K-nearest neighbors
- Redes Neurales
- Decision Trees



No Supervisados

- K-Means
- Métodos jerárquicos
- Mezcla Gaussiana
- DBScan
- Mini-Batch-K-Means



¡Estamos aquí!

Clasificación de los algoritmos



Respuesta continua

- Regresión
- GLM
- GLMM
- Support vector machines
- K-nearest neighbors
- Redes Neurales
- Regression Trees




Respuesta discreta

- Regresión logística
- Clasification trees
- Redes Neurales
- GLM
- GLMM


¡Estamos aquí!

Clasificación de los algoritmos



Métodos Machinelárnicos

- Regresión
- GLM
- GLMM
- ANOVA



Métodos Machinelárnico- estadísticos

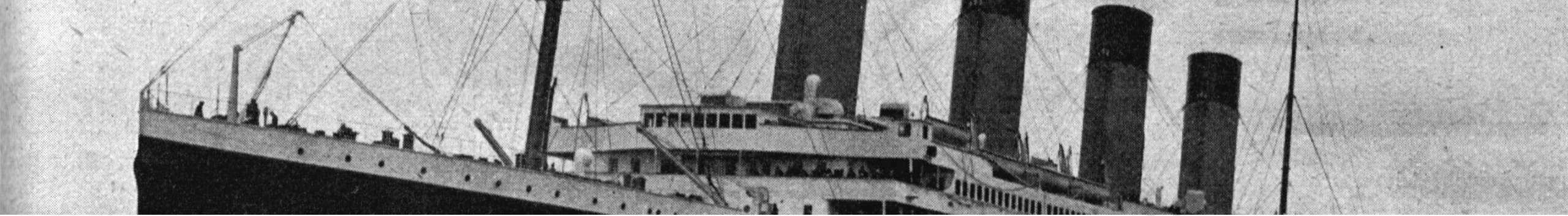
- Árboles de decisión
- Bagging
- Boosting
- K-NN
- Redes Neurales
- Support vector machines

¡Estamos aquí!



Nuestro problema: clasificar sobrevivientes

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



Reflexiones sobre la base de datos

Población

- ~ 2.200 personas
- ~ 1.300 pasajeros
- Más de 1.500 muertos

Muestra

- 891 personas
- 549 no sobrevivientes
- 342 sobrevivientes

Objetivos del algoritmo

- Clasificar de la mejor forma posible la variable respuesta
 - ... A través de segmentaciones
 - ... Usando las variables explicativas
- Obtener insights
 - ... De las relaciones entre la variable respuesta y las explicativas
 - ... Explorar interacciones

Luiz Rodriguez Fantini 005.374.619-81

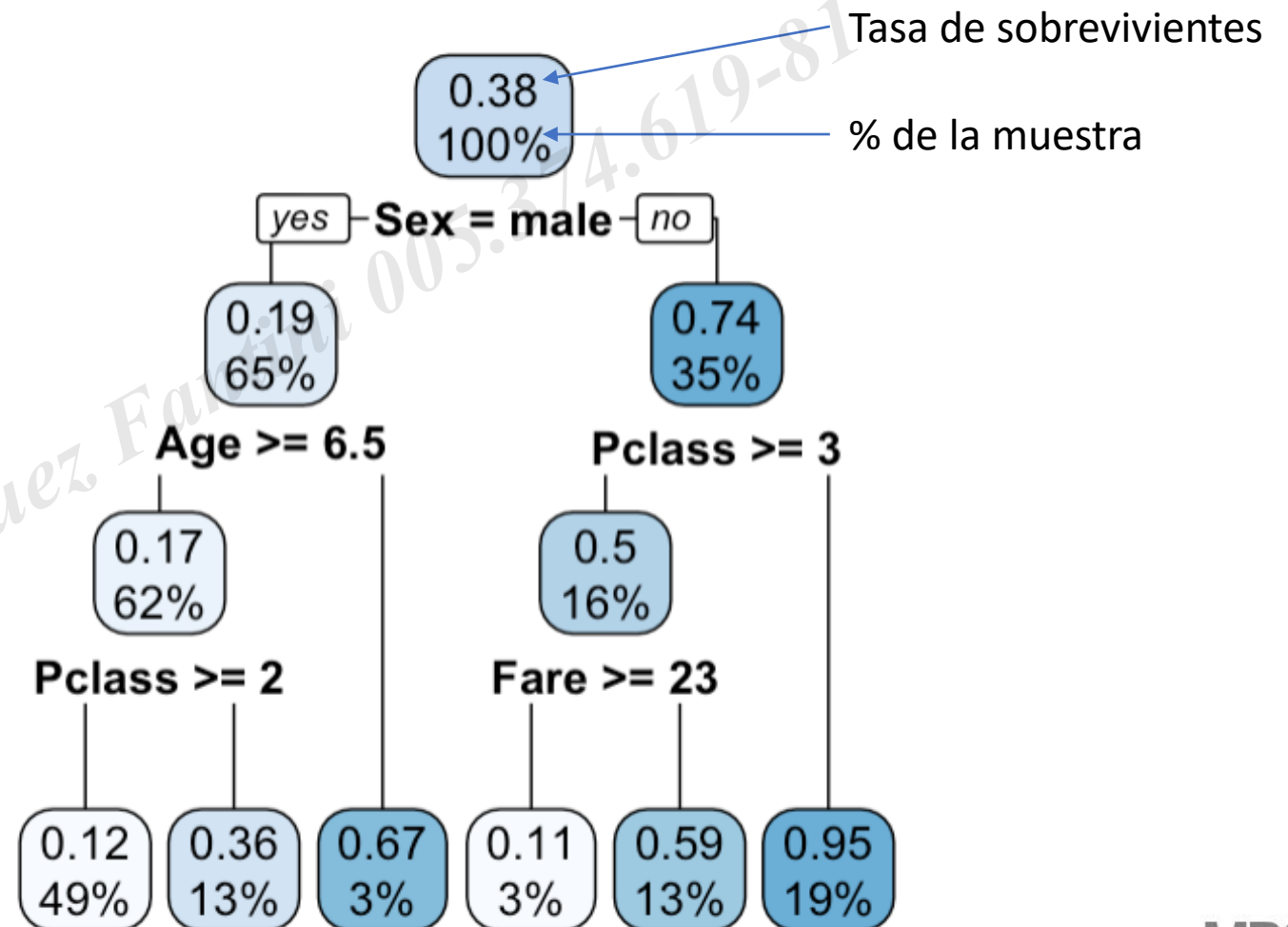


OMML1_script01-Primeiro_contato_com_arvores.R

¿Qué es un árbol de decisión?

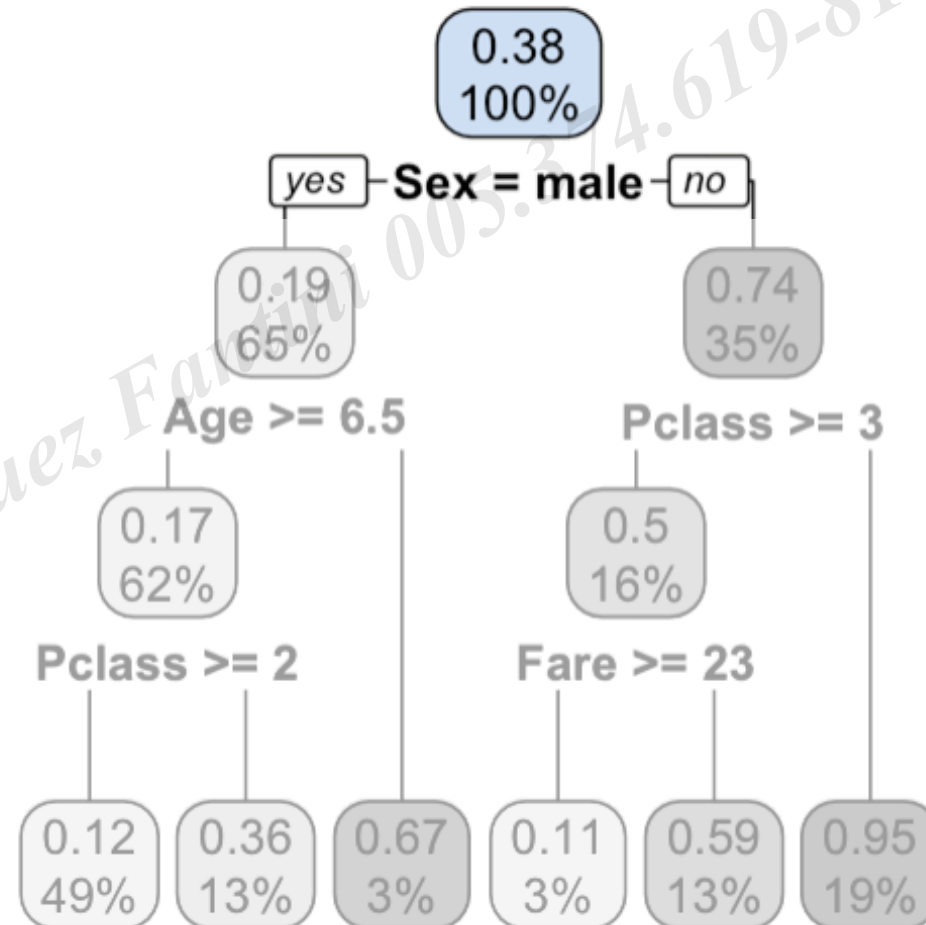
El árbol de decisión es:

Una secuencia de segmentaciones binarias
Que pretende homogeneidad de la variable
respuesta



¿Qué es un árbol de decisión?

Inicialmente tenemos 891 pasajeros de los cuales
342 sobrevivieron (38%)
549 no sobrevivieron

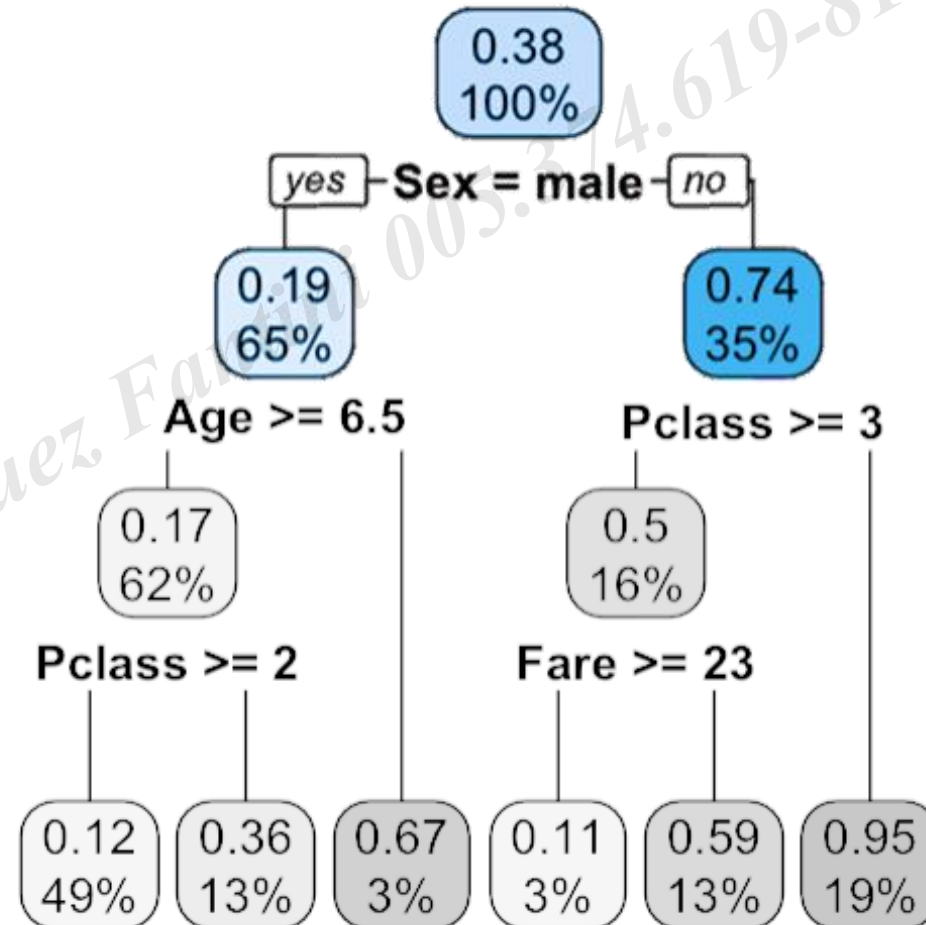


¿Qué es un árbol de decisión?

De los 891, podemos segmentarlos en:

577 hombres (65%) de los cuales
109 sobrevivieron (19%)
468 no sobrevivieron

314 mujeres (35%) de las cuales
233 sobrevivieron (74%)
81 no sobrevivieron



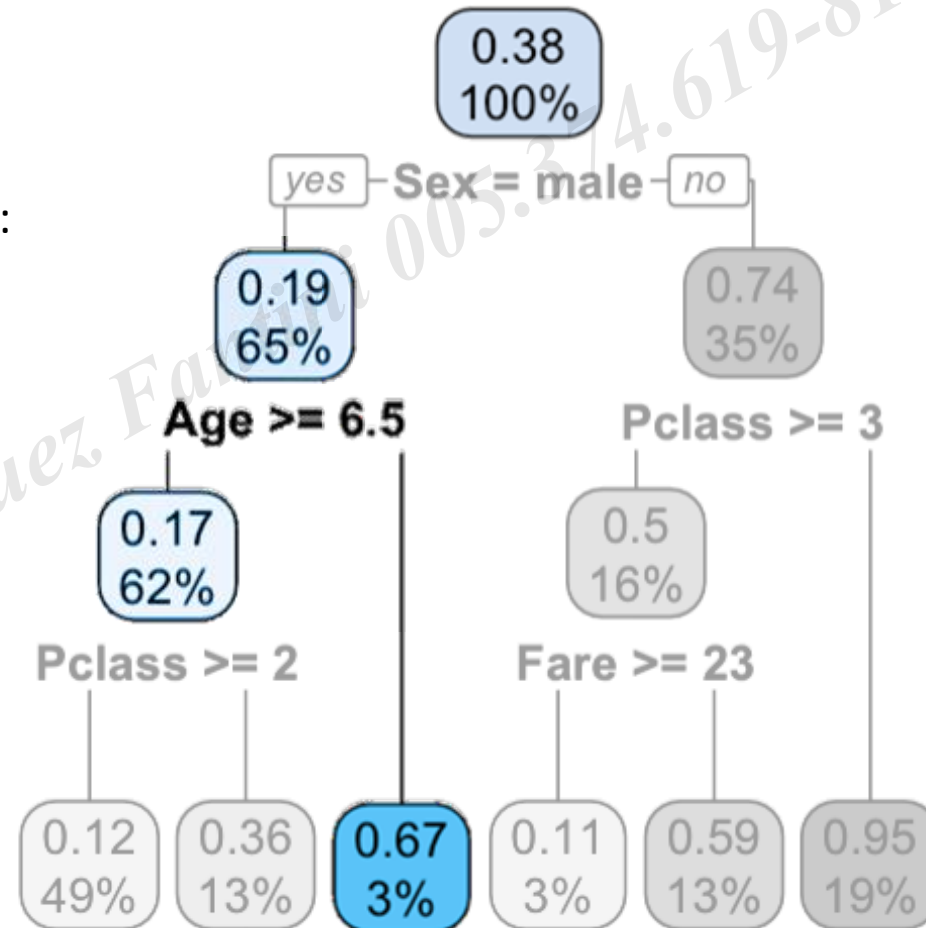
¿Qué es un árbol de decisión?

De los 891, podemos segmentarlos en:

577 hombres que por su vez segmentamos en:

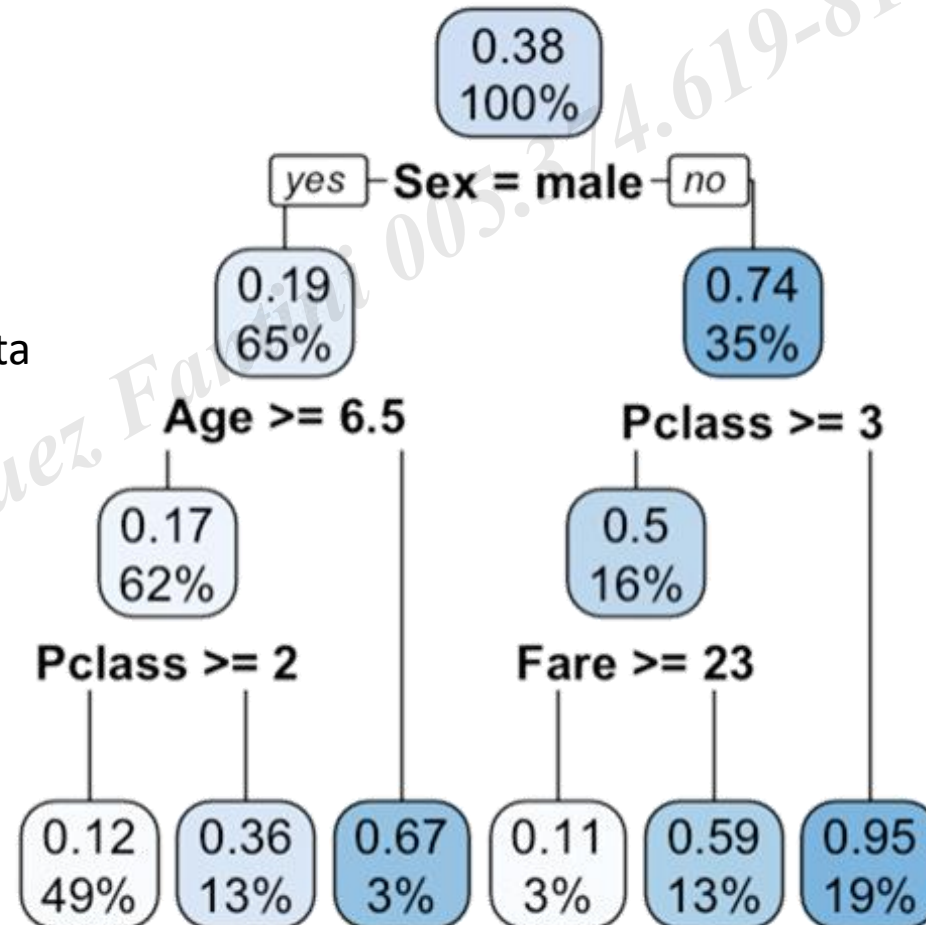
24 niños (< 6,5 años) de los cuales
16 sobrevivieron (67%)
8 no sobrevivieron

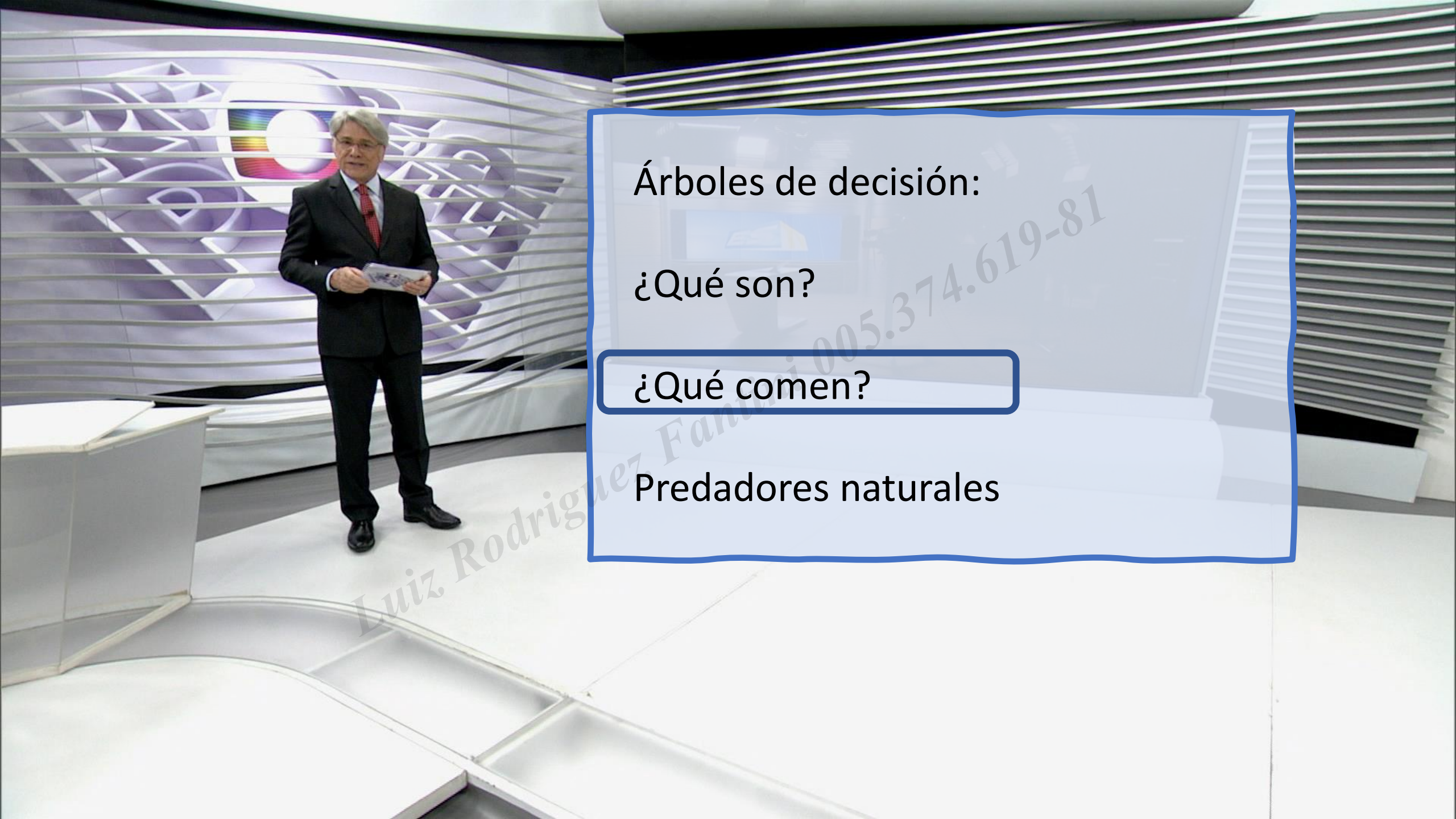
533 adultos ($\geq 6,5$ años) de los cuales
93 sobrevivieron (17%)
553 no sobrevivieron



¿Qué es un árbol de decisión?

Y así continuamos "meneando" la muestra hasta que "no valga la pena" hacer más quiebras.





Árboles de decisión:

¿Qué son?

¿Qué comen?

Predadores naturales

Definiciones de impureza

- Gini

- Entropía de Shannon

¿Cómo el árbol encuentra el mejor quiebre?
Con una métrica de 'impureza'

Índice de Gini

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

- Impureza máxima con distribución uniforme
- Impureza mínima en la concentración total

Entropia

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

Ganancia de información:

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

- Impureza máxima con distribución uniforme
- Impureza mínima en la concentración total

Algoritmo básico

1. Para cada variable, buscar la mejor regla binaria
2. Elegir aplicar mejor segmentación entre todas las variables
3. Recursivamente, para cada hoja, repetir los pasos 1 y 2 hasta que una regla de parada sea alcanzada

Implementación web interactiva:

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

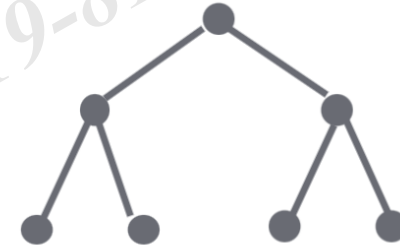
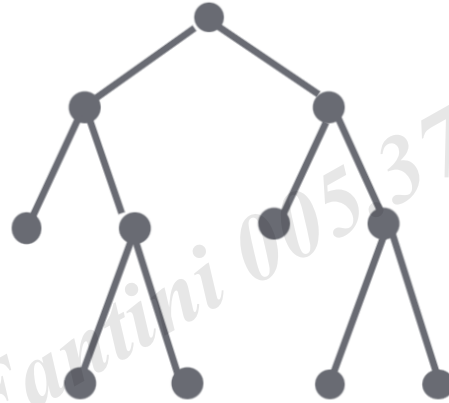
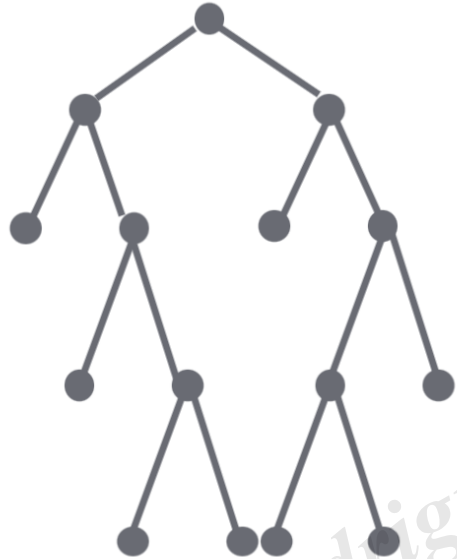
Hiperparámetros

Son parámetros que controlan el algoritmo como:

1. Número mínimo de observaciones por hoja
2. Profundidad máxima
3. CP – Costo de complejidad

Luiz Rodriguez Fantini 005.374.619-81

Costo de complejidad



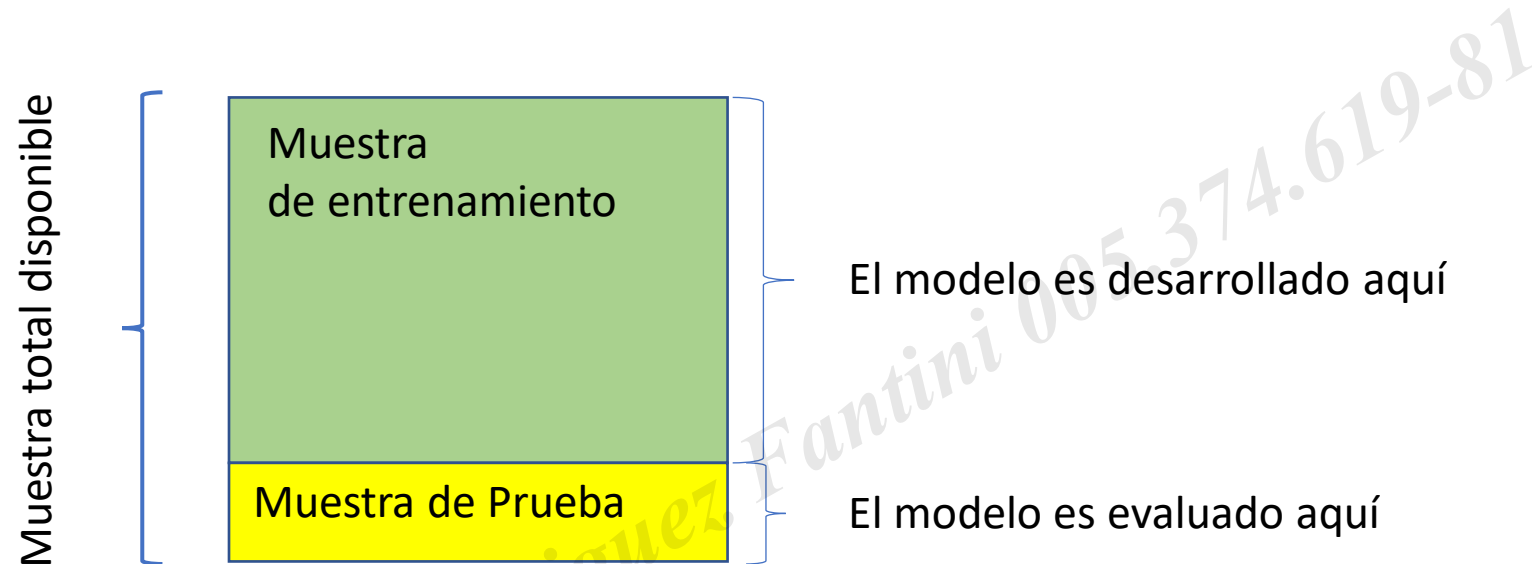
Costo de
complejidad

Bajo

Medio

Alto

Cross validation (validación cruzada)



La estrategia más simple es dividir la base en entrenamiento y prueba.
Desarrollamos el modelo en la base de entrenamiento y evaluamos en la base de prueba.

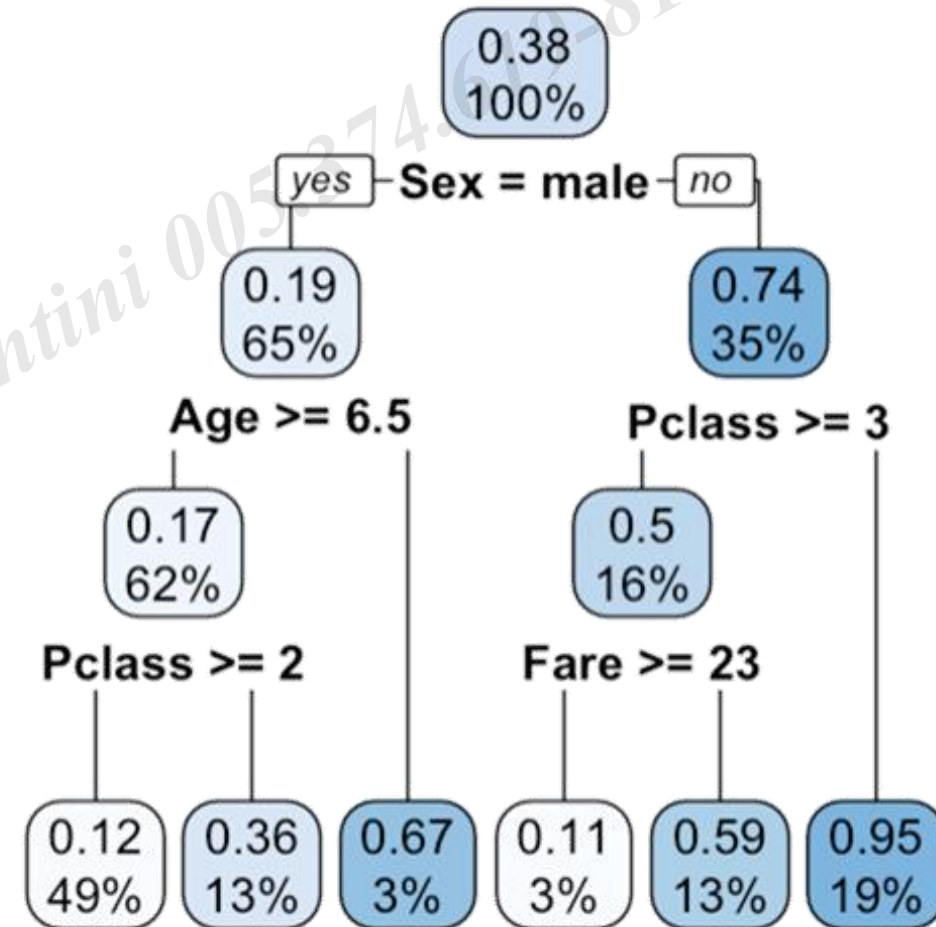


OMML1 _script02-Algoritmo_avaliacao_overfitting

El árbol como un clasificador

Requisitos:

Tener todas las variables.



El árbol como un clasificador

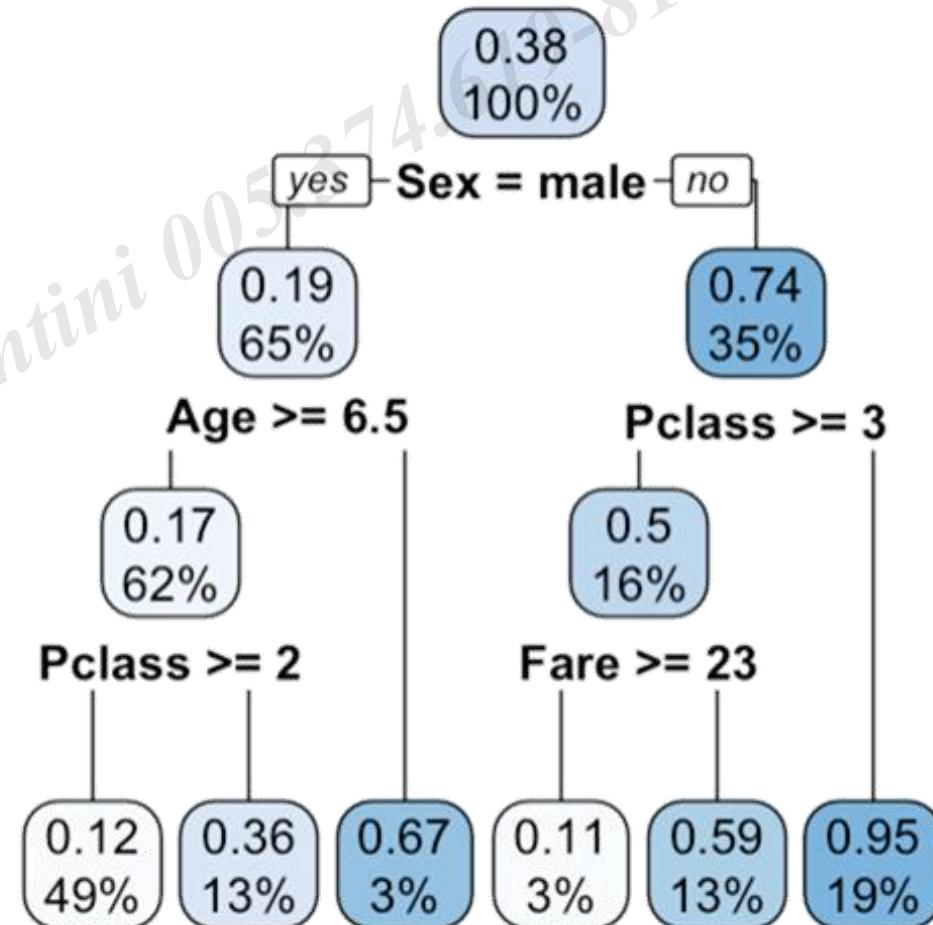
Probabilidad de evento de la hoja F:

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

$P(S|F)$ - probabilidad de éxito de la hoja F

N_f - es el número de individuos en la hoja F

N_f^S - es el número de sobrevivientes en la hoja F



El árbol como un clasificador

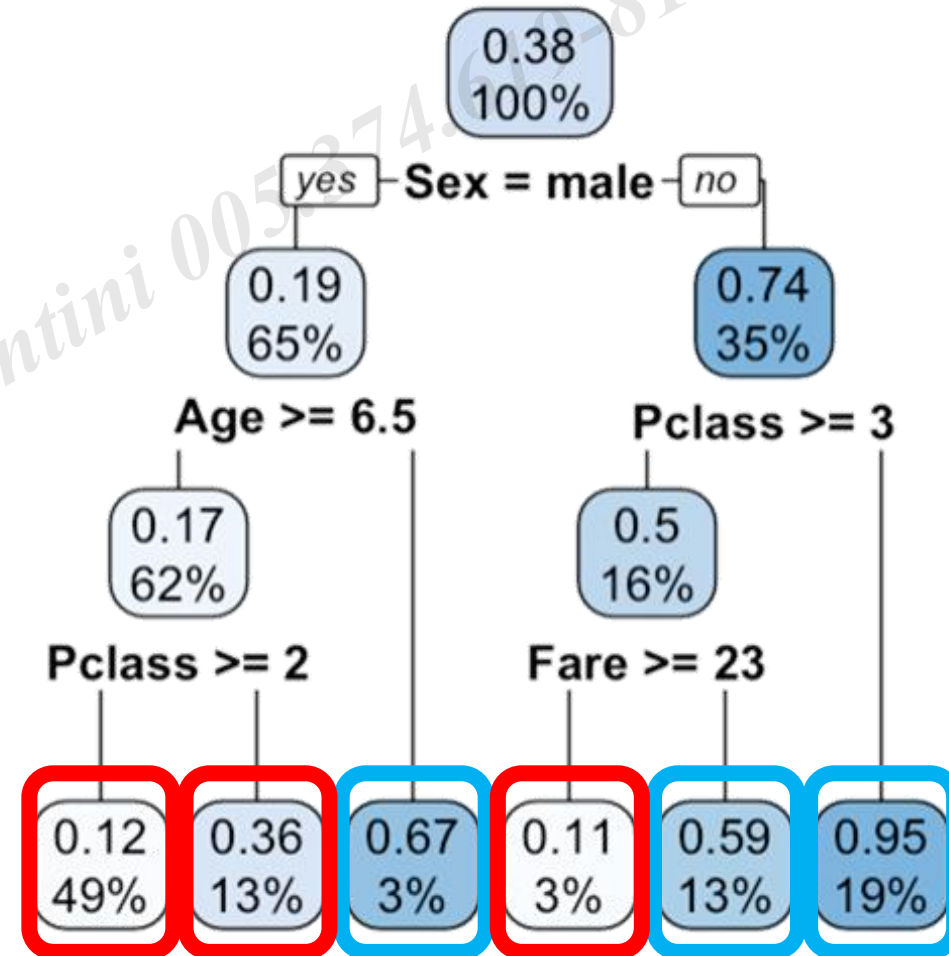
Clasificación:

Clasificación estándar:

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

No sobrevivientes: $P(S|F) < 50\% \Rightarrow C(F) = "N"$

Valor predicho	Valor Verdadero	
	0	1
0	484	96
1	65	246





Evaluación del modelo

- Exactitud:

Aciertos sobre intentos

Valor predicho	Valor Verdadero	
	0	1
0	484	96
1	65	246

En el ejemplo:

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

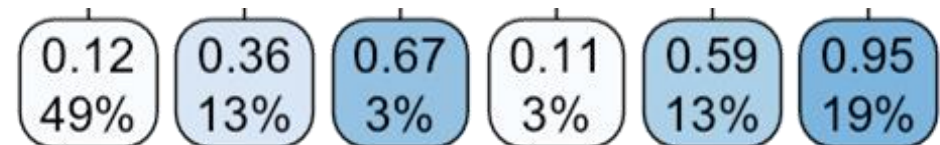
Árbol como diagnóstico

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

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

Valor predicho	Valor Verdadero	
	0	1
0	484	96
1	65	246

Valor predicho	Valor Verdadero	
	0	1
0	TN	FN
1	FP	TP



Diagnóstico y puntos 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

Exactitud	Especificidad	1- Especificidad	Sensibilidad
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 punto de corte, tenemos una matriz de confusión.
En el caso, tenemos 8 posibles matrices con el árbol entrenado.

Curva ROC

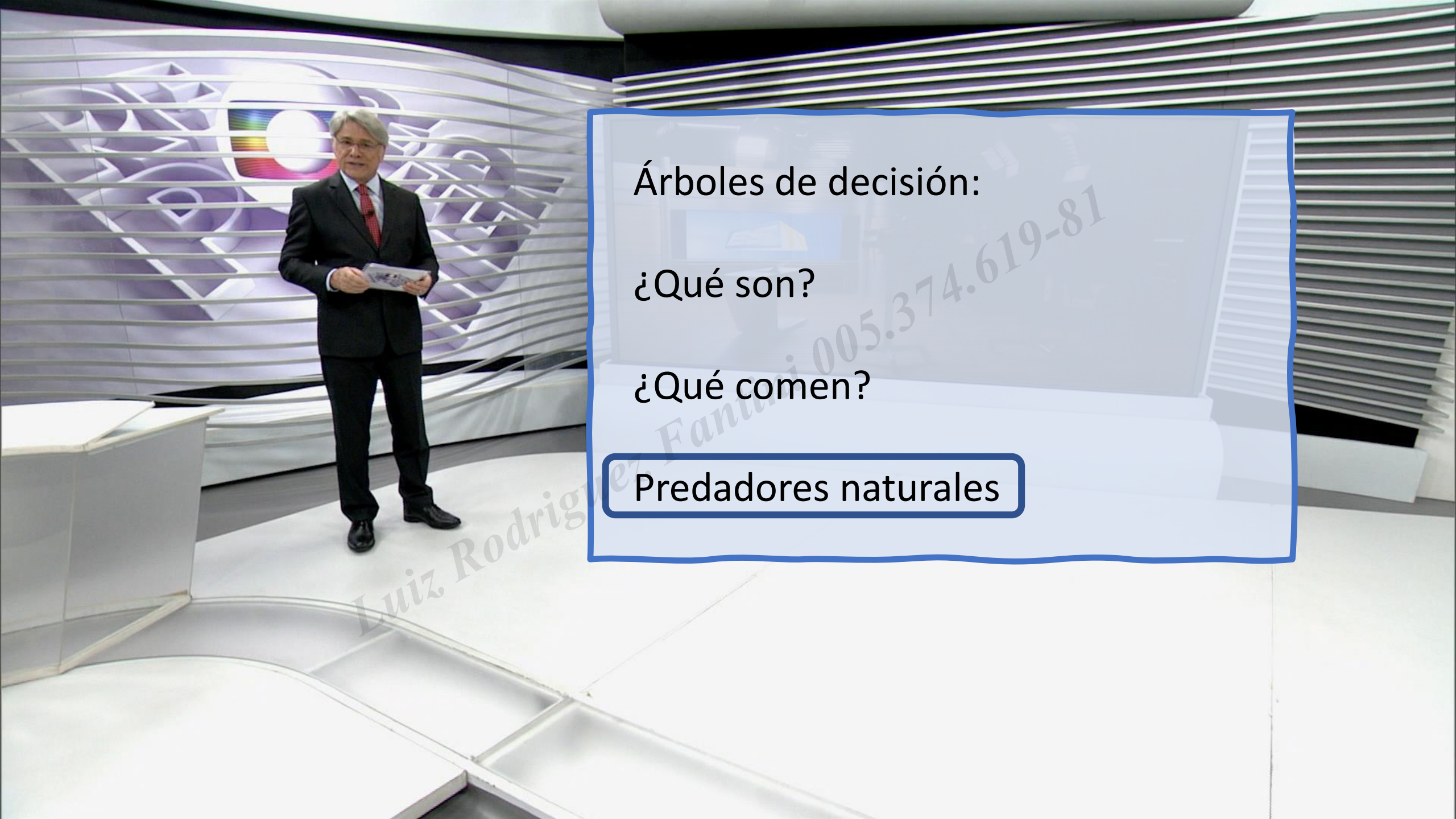
Corte	1- Especificidad	Sensibilidad
0% - 11,1%	100%	100%
11,1% - 11,5%	96%	99%
11,5% - 35,8%	26%	85%
35,8% - 58,9%	12%	72%
58,9% - 66,7%	3%	52%
66,7% - 94,7%	2%	47%
94,7% - 100%	0%	0%



La curva ROC es un gráfico de dispersión de 1-Especificidad en el eje X por Sensibilidad en el eje Y, obtenidos para cada posible punto de corte del clasificador.



OMML1 _script02-Algoritmo_avaliao_overfitting



Árboles de decisión:

¿Qué son?

¿Qué comen?

Predadores naturales

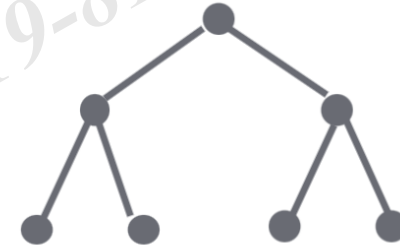
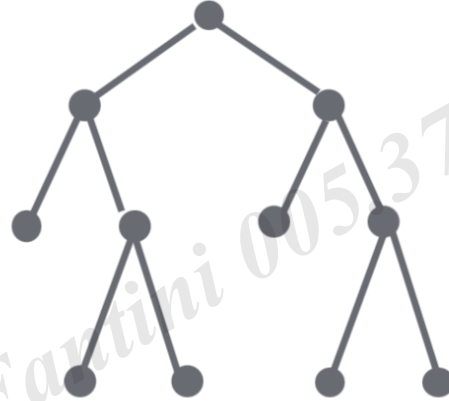
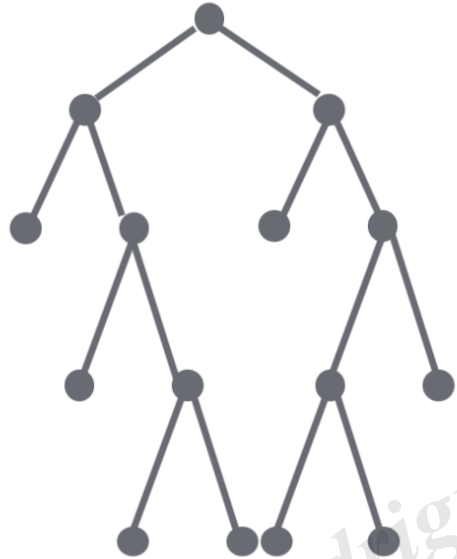
A photograph of a bed frame with a dark wood headboard and footboard. On the bed is a white, quilted mattress that has been shaped into a highly complex, convoluted form, resembling a tangled path or a highly specific, non-generalizable model. The bed is set on a light-colored wooden floor.

THE BEST WAY TO EXPLAIN OVERFITTING

Qué es

Cómo evitar

Poda del árbol (*Prunning*)



EXACTITUD

Base de entrenamiento: 95%
Base de validación: 40%

Base de entrenamiento: 70%

Base de validación: 60%

Base de entrenamiento: 65%

Base de validación: 64%

MUESTRA DE ENTRENAMIENTO

MUESTRA DE VALIDACIÓN

Estrategias de cross validation

Elegir parámetros del modelo con una base de validación puede generar overfitting

Existen diversas técnicas de validación cruzada para evitar ese efecto. En este momento voy a mencionar una técnica clásica: dividir la base en Entrenamiento, Validación y Prueba



Muestra de entrenamiento

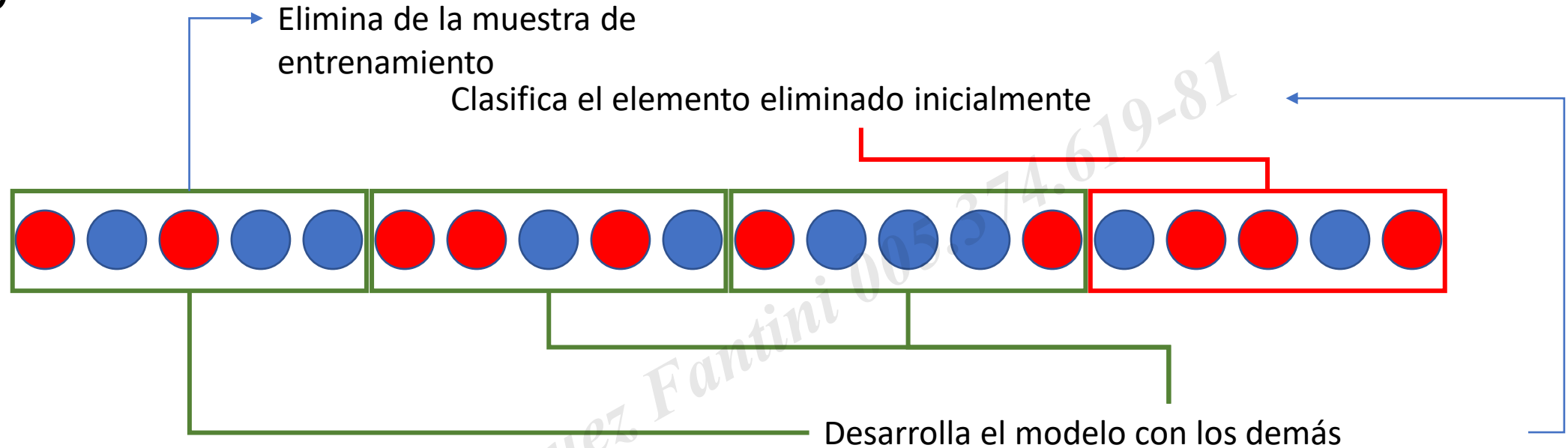


Muestra de validación



Muestra de prueba

K-fold



- Dividimos la base en k submuestras
- Para cada submuestra:
 - Eliminamos la submuestra como validación
 - Entrenamos el modelo con las observaciones restantes
 - Utilizamos este modelo para clasificar la submuestra eliminada
 - Evaluamos la métrica de desempeño del modelo
- Calculamos la media de las métricas de desempeño del modelo

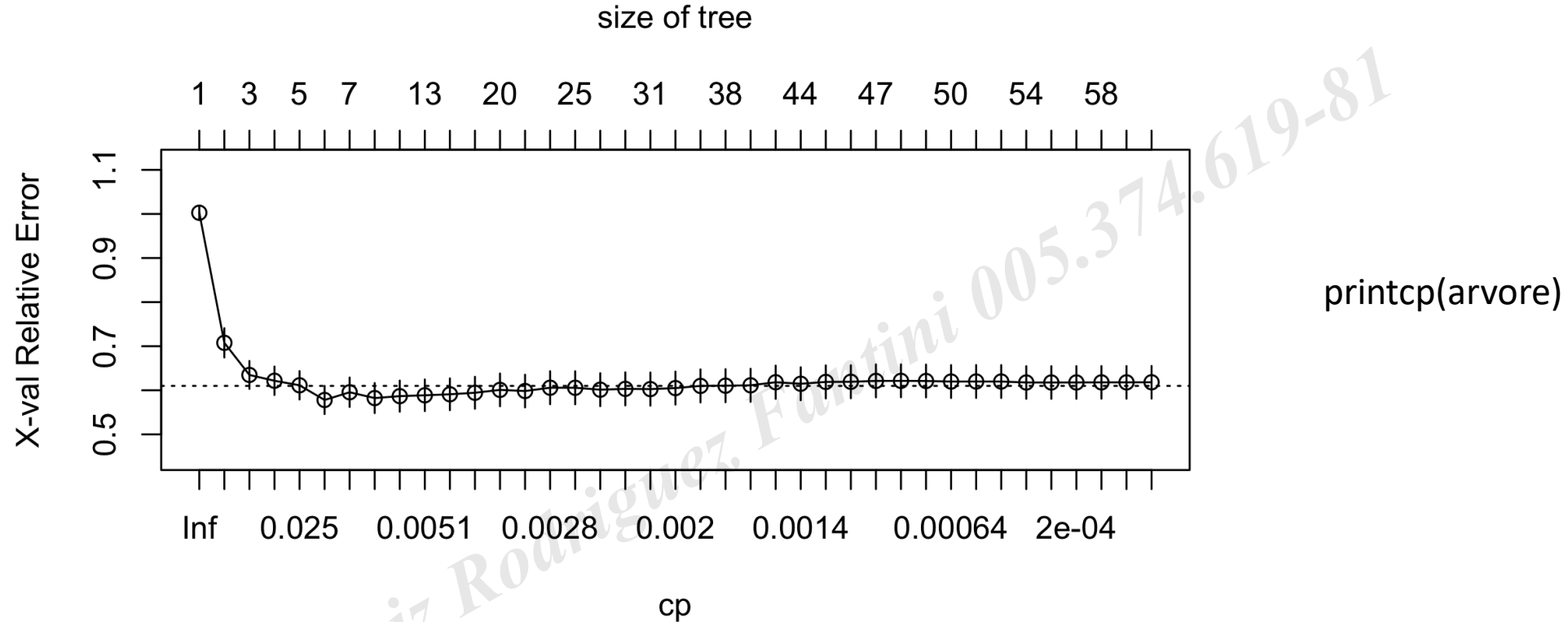
K-fold

Típicamente, hacemos lo mismo para variaciones del modelo para optimizar hiperparámetros.



	Exactitud 1	Exactitud 2	Exactitud 3	Exactitud 4	Exactitud Media
Modelo 1	62%	58%	61%	59%	60%
Modelo 2	50%	51%	49%	47%	49%
Modelo 3	72%	68%	71%	75%	72%

Post-prunning con crossvalidation



R hace la poda del árbol realizando un k -fold para optimizar el CP (complexity path), un parámetro que sintetiza la complejidad del árbol. Eso es realizado con un k -fold.



OMML1 _script02-Algoritmo_avaliao_overfitting

A photograph of two skiers on a snowy mountain peak. The skier in the foreground is wearing a red jacket and orange pants, standing on a snowdrift. The second skier is further back, wearing a blue jacket and dark pants, also on the snow. The background shows more snow-covered mountain ridges under a bright blue sky with scattered white clouds.

Conclusión

- Robustas, interpretables, flexibles
- Sin suposiciones probabilísticas
- Necesario *cross-validation*

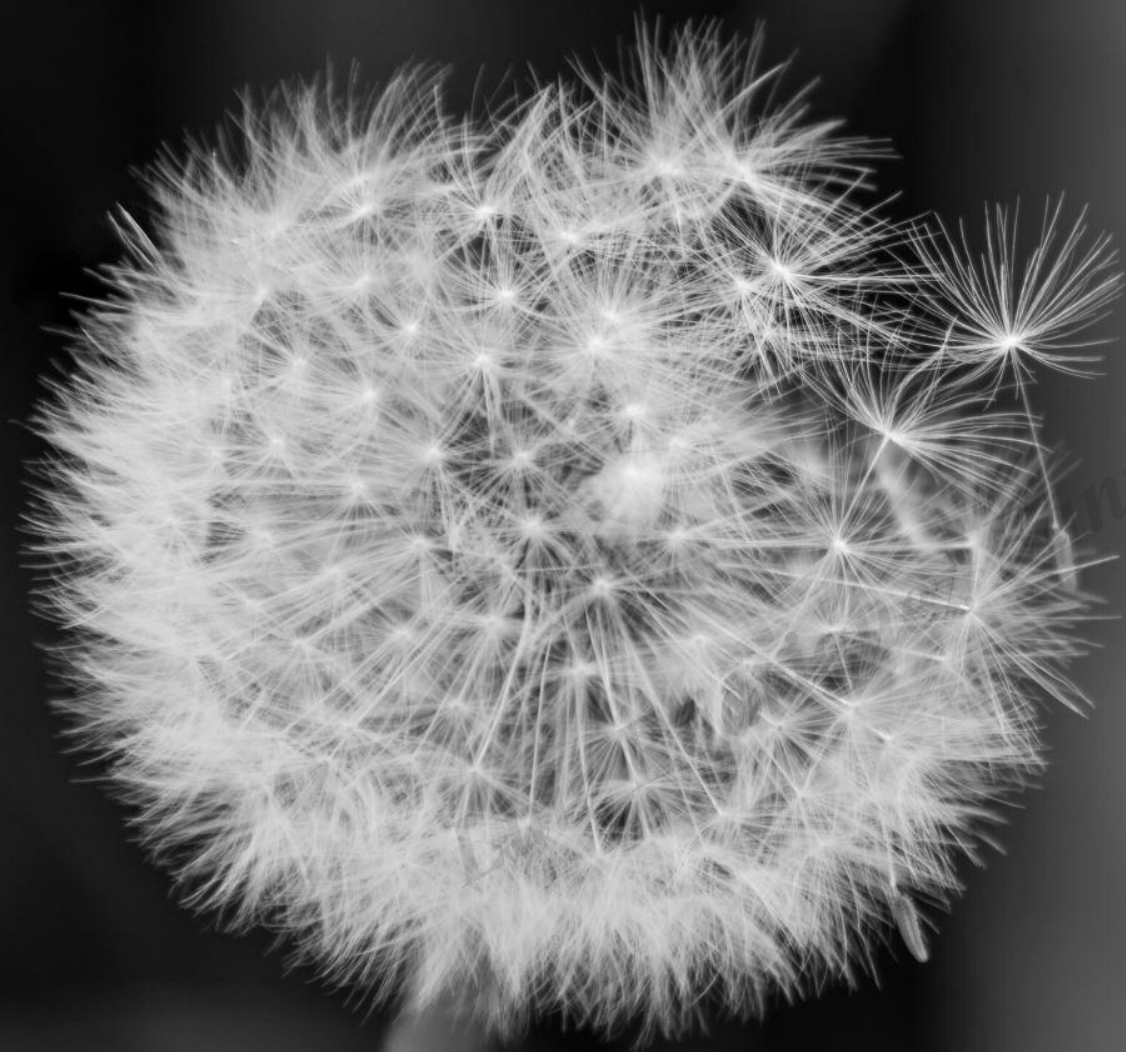
Quanto mais aprendo, mais
tenho certeza de que, o que
sei, é apenas uma gota,
diante do oceano do que
ainda preciso aprender.



PENSADOR

Jose Ap Barcelos

Cuánto más aprendo, más estoy seguro de que, lo que sé, es apenas una gota,
frente al océano de lo que todavía preciso aprender.



Por hoy es sólo
eso ;)



[linkedin.com/in/joao-serrajordia](https://www.linkedin.com/in/joao-serrajordia)

Algoritmos famosos

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

Luiz Rodriguez Fantini 005.374.619-81

Stack overflow interesante sobre eso.

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