esden
Business School

R

Leonardo Hansa

Agenda (variable)

1. Aterrizaje
2. Modelos de regresión
   - Modelos lineales (ejemplo longitudinal).
   - *Modelos basados en árboles* (ejemplo transversal).
3. Modelos de clasificación
   - Regresión logística.
   - *SVM*.
4. Modelos no supervisados
   - Algoritmo kmeans.
   - Algoritmo Clara.
5. Siguientes pasos
   - Problemas con la regresión.
   - Redes neuronales.
   - Clusters jerárquicos.

Aterrizaje

¿Quién soy?

- Leonardo Hansa
- Matemáticas (UCM)

¿Qué hago?

- Data scientist y data analyst:
  - Conento (Deloitte)
  - DIA
  - Minsait (Indra)
  - Ebiquity
- Comunidad R Hispano

¿Dónde estoy?

hola@leonardohansa.com

La realidad cambia

Aterrizaje

¿Qué es un modelo?

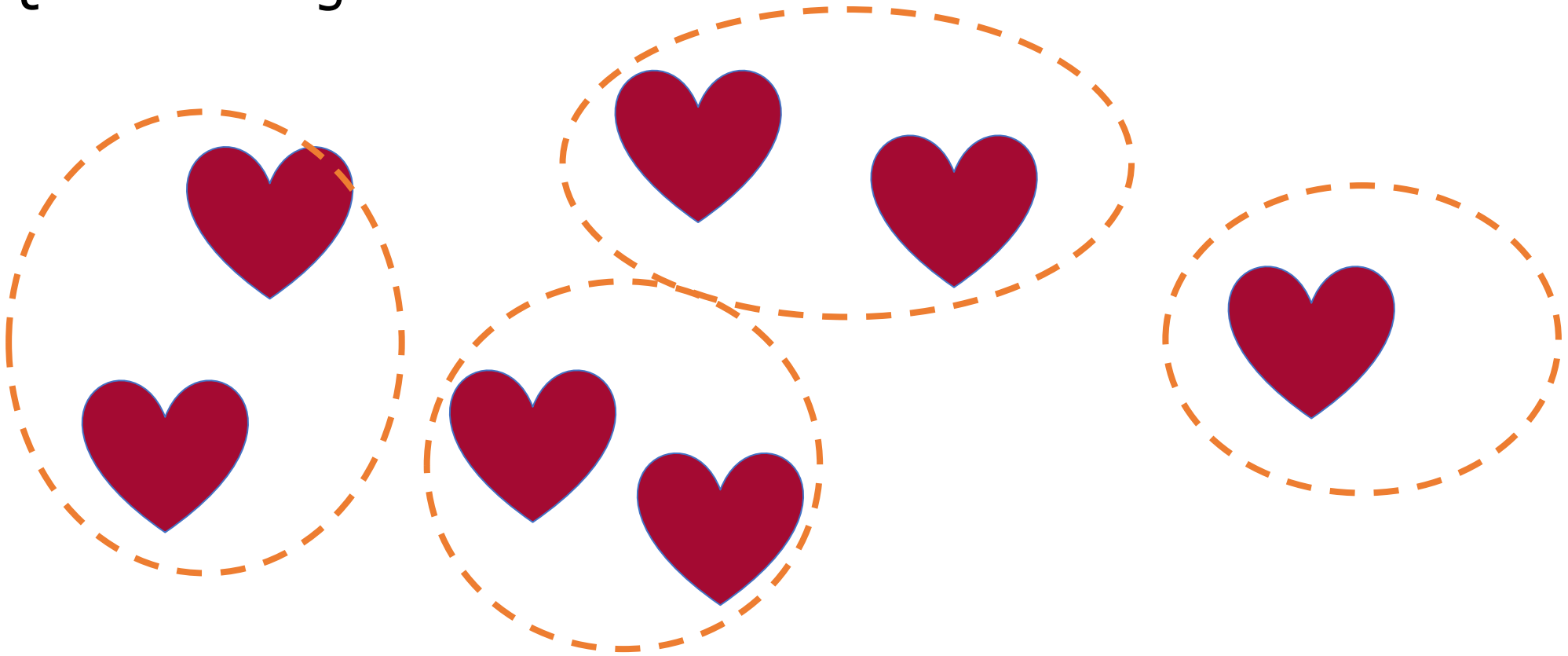
Aterrizaje

¿Qué es un modelo matemático?

$$y = \alpha + \beta \cdot x$$

Aterrizaje

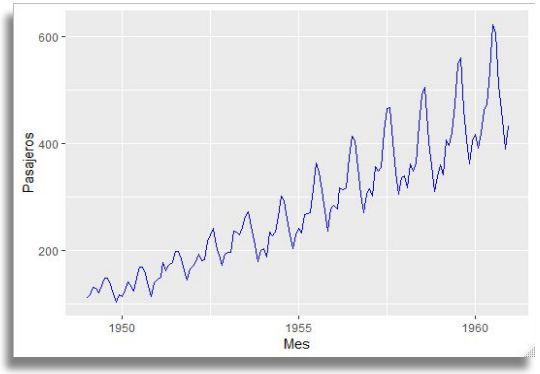
¿Qué es un algoritmo?



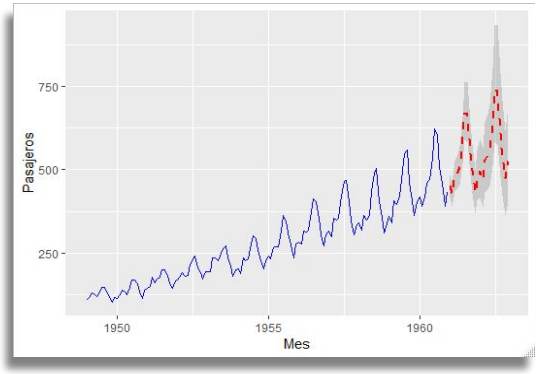
Aterrizaje

Sin histórico, no hay modelos

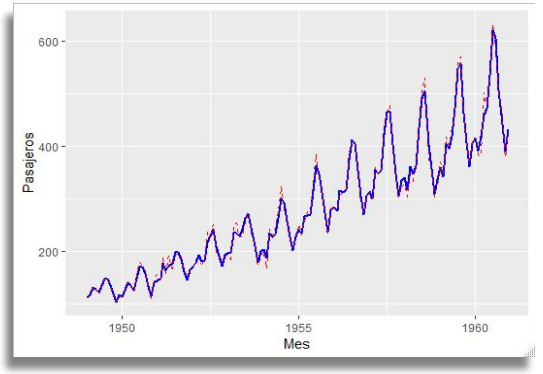
Paso 1. Datos históricos.



Paso 2. Ajustamos un modelo que aproxime el histórico.

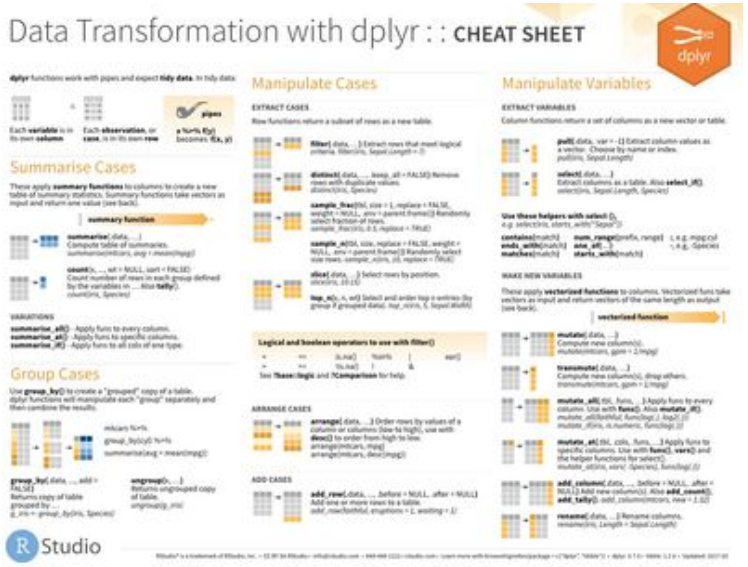


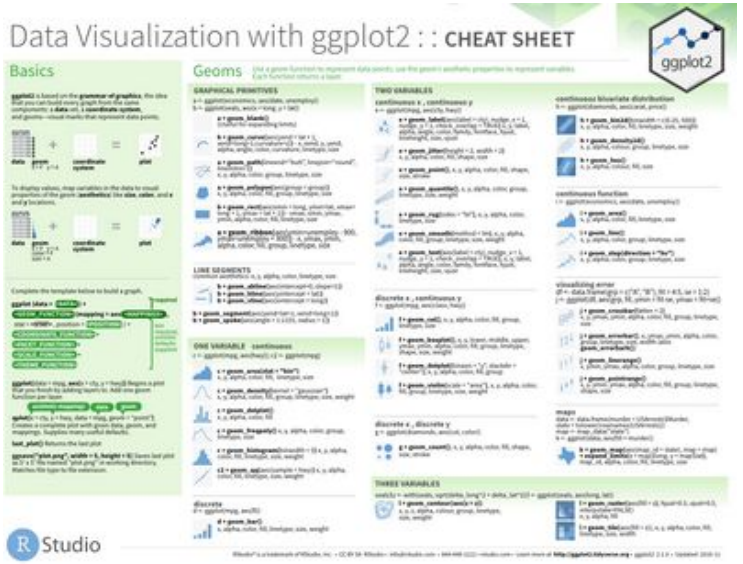
Paso 3. Extrapolamos hacia el futuro.



Aterrizaje

https://www.rstudio.com/resources/cheatsheets/

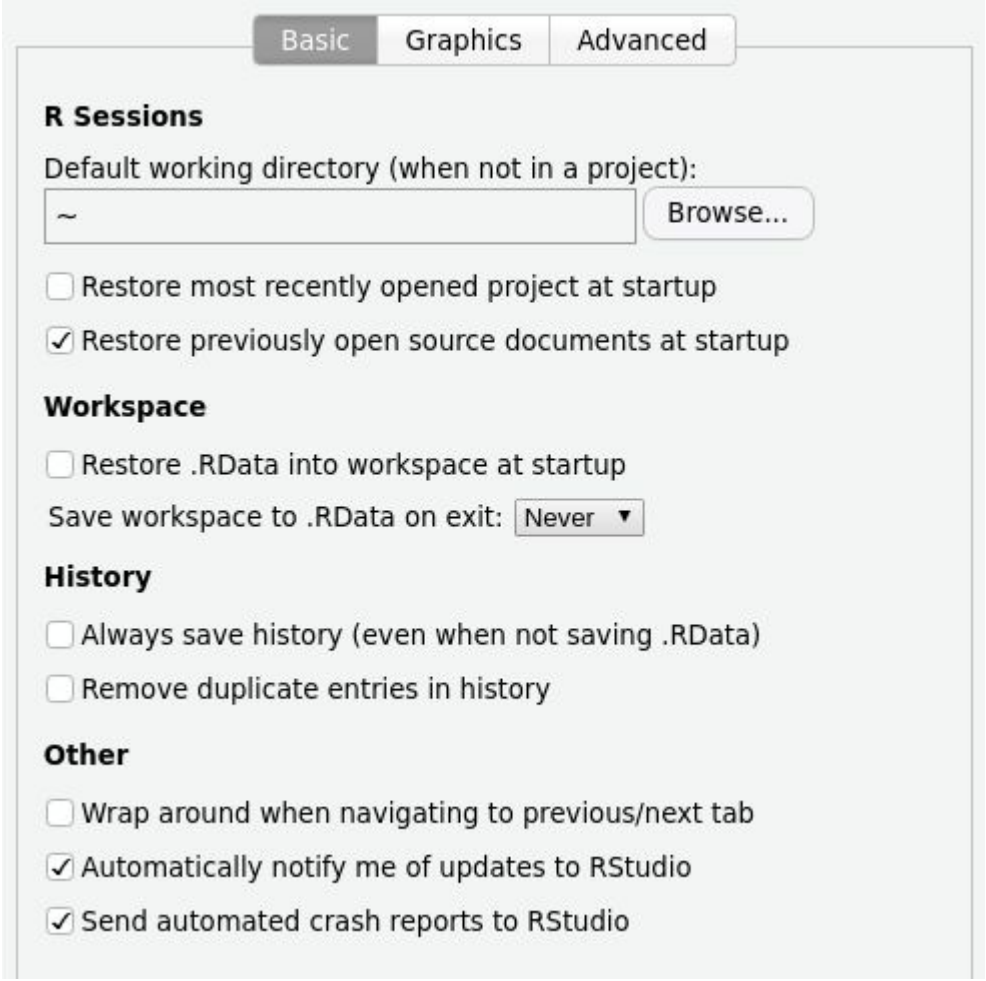




Aterrizaje

RStudio®

Tools > Global options



esden
Business School

Regresión

Cuando modelamos datos continuos

Regresión

Regresión lineal

Coeficientes

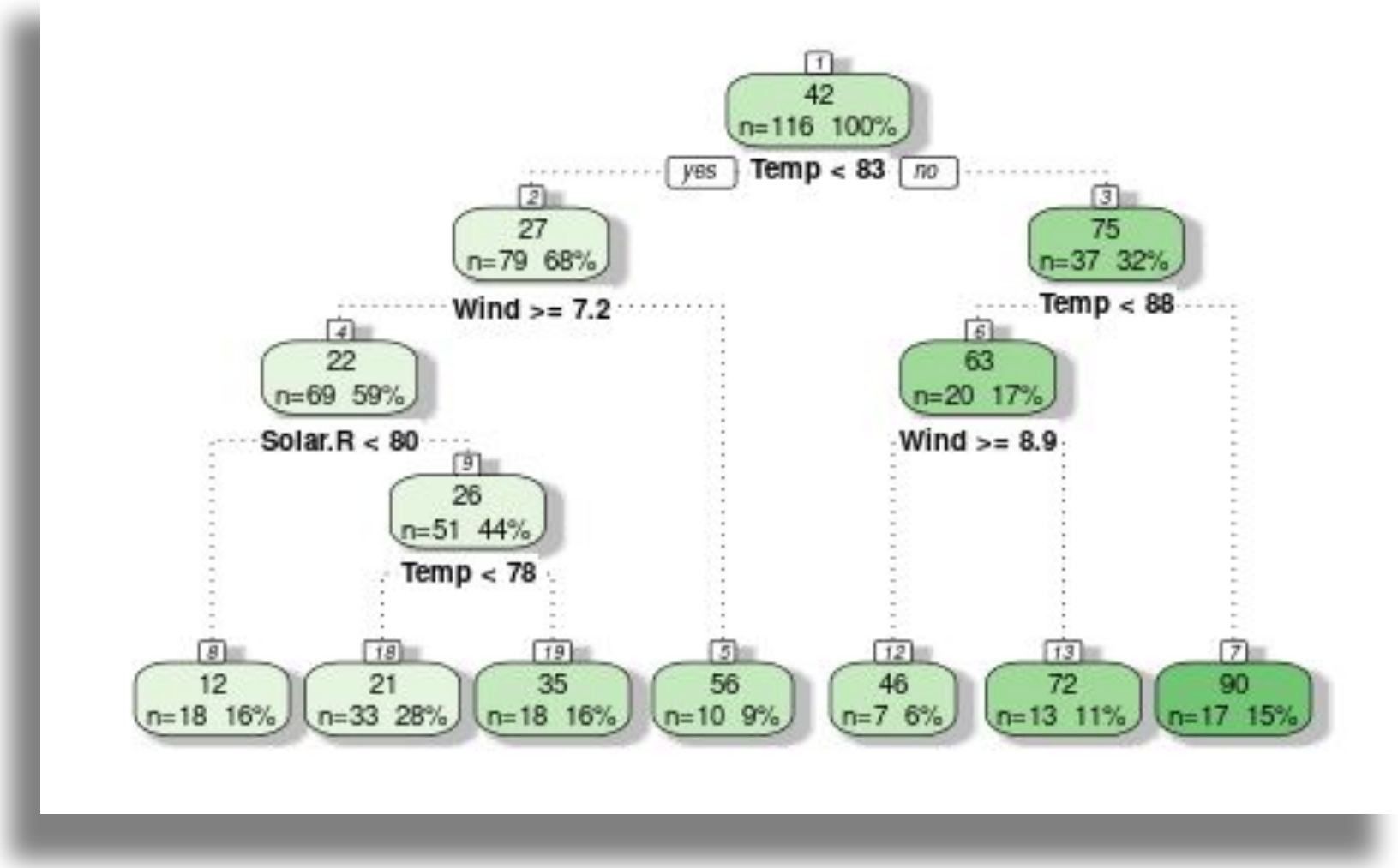
P-valores

R2 y R2 ajustado

```
Call:
lm(formula = ventas ~ +1 + p2ola1_ad40 + p3_ola1_ad40 + p4_ola1_ad40 +
    p5_ola1_ad40 + p1_ola2_ad40 + p1_ola3_ad40 + dp + competencia1 +
    competencia2, data = df_consumo)

Residuals:
     Min       1Q   Median       3Q      Max
-1.14202 -0.34507 -0.09212  0.25147  1.87867

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.2851497  0.4408759 -11.988  < 2e-16 ***
p2ola1_ad40  0.0078508  0.0013079   6.003 1.43e-08 ***
p3_ola1_ad40 0.0039914  0.0005200   7.676 2.05e-12 ***
p4_ola1_ad40 0.0050948  0.0008755   5.819 3.53e-08 ***
p5_ola1_ad40 0.0057182  0.0008666   6.598 6.94e-10 ***
p1_ola2_ad40 0.0045156  0.0011075   4.077 7.41e-05 ***
p1_ola3_ad40 0.0029815  0.0009479   3.145  0.00201 **
dp           0.2617109  0.0110016  23.789  < 2e-16 ***
competencia1 -0.0027755  0.0013679  -2.029  0.04424 *
competencia2 -0.0031153  0.0011652  -2.674  0.00835 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5353 on 148 degrees of freedom
Multiple R-squared:  0.8939,    Adjusted R-squared:  0.8874
F-statistic: 138.5 on 9 and 148 DF,  p-value: < 2.2e-16
```

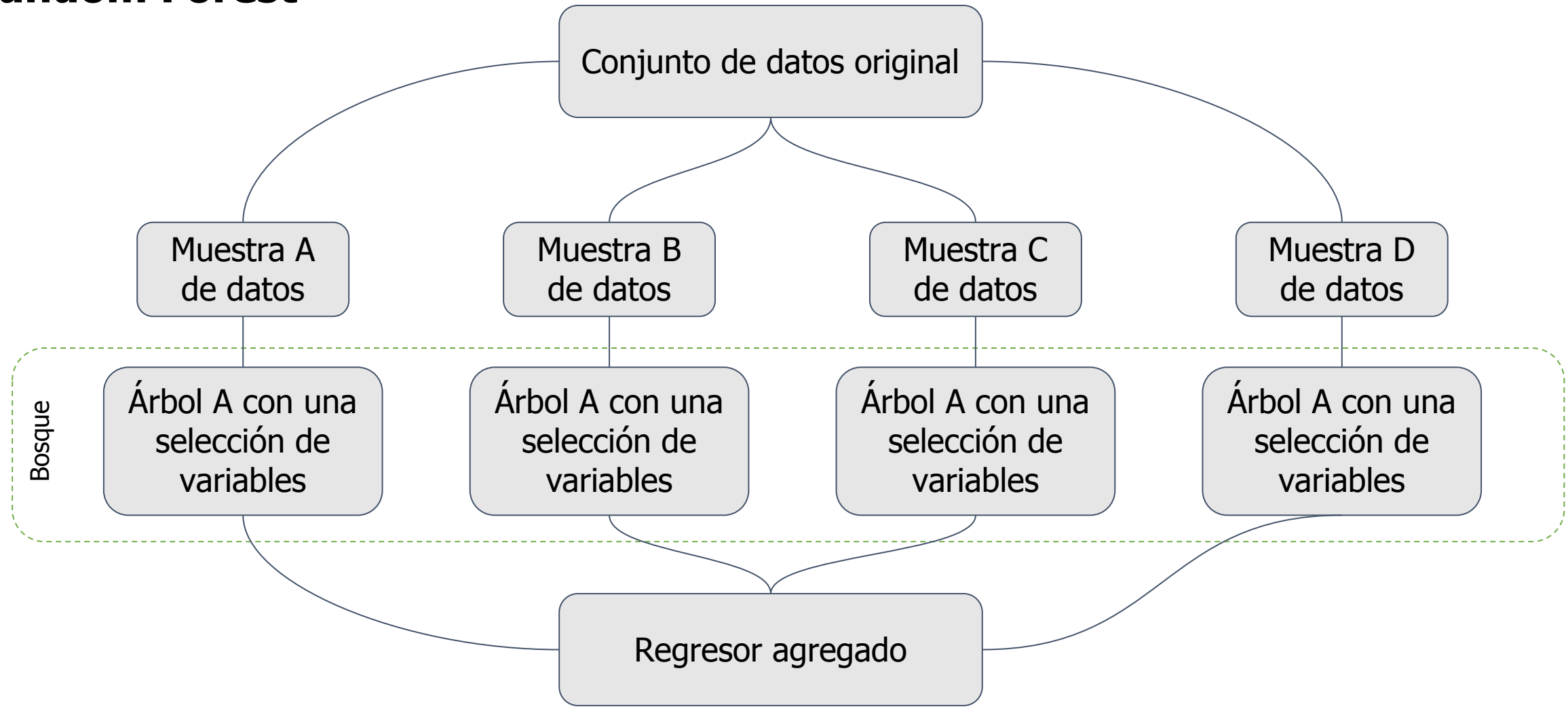
Regresión

Árbol de regresión



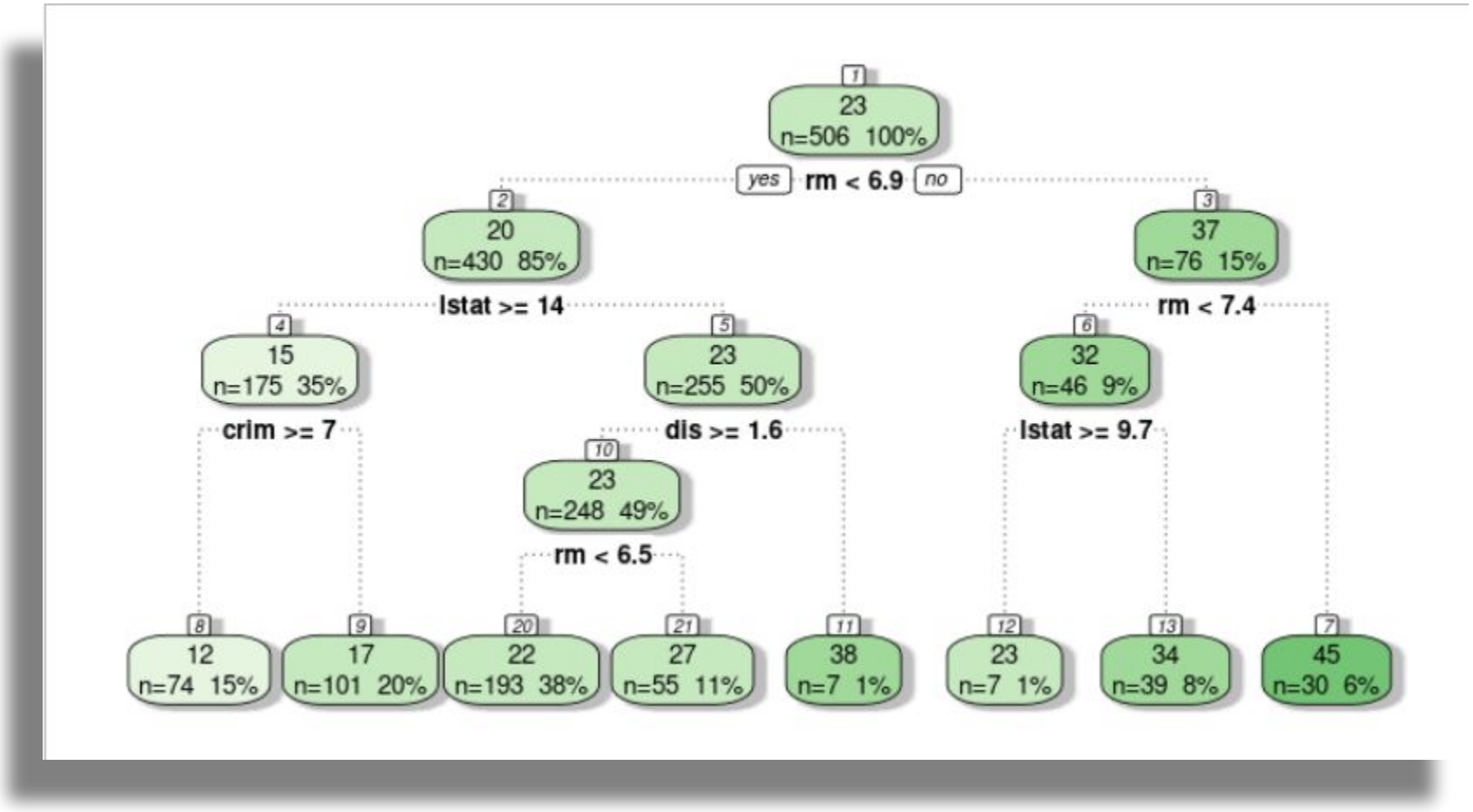
Regresión

Random Forest



Regresión

Boston



esden
Business School

Clasificación

Cuando modelamos categorías

Clasificación

Framingham

ten_year_chd.
padeció enfermadad cardiovascular en los 10 años siguientes al examen.

- male.
  - 0: mujeres; 1: para varones.
- age.
  - Edad en el momento del examen médico.
- high_school_ged.
  - 1: graduado escolar como máximo nivel de estudios.
- some_college_vocational_school.
  - 1: diplomado.
- college.
  - 1: licenciado.
- current_smoker.
  - 1: fumador; 0: no fumador.
- cigs_per_day.
  - Número de cigarrilos al día (media estimada).
- bp_meds.
  - 0: Sin medicamentos por tensión alta; 1: con medicación contra tensión alta.
- prevalent_stroke.
  - 1: riesgo de derrame.
- prevalent_hyp.
  - 1: riesgo de hipertensión.
- diabetes.
  - 1: diabetes.
- tot_chol.
  - Colesterol en mg/dL.
- sys_bp.
  - Presión arterial sistólica.
- dia_bp.
  - Presión arterial diastólica.
- bmi.
  - Índice de masa corporal.
- heart_rate.
  - Ritmo cardíaco.
- glucose.
  - Índice de glucosa en sangre.

Clasificación

Matriz de confusión

	Real: 0	Real: 1	
Predicción: 0	1.989	204	Falsos negativos: 0,093
Predicción: 1	1.112	353	Verdaderos positivos: 0,24
	Especificidad: 0,6414	**Sensibilidad: 0,6338**	**Precisión: 0,6402**

Clasificación

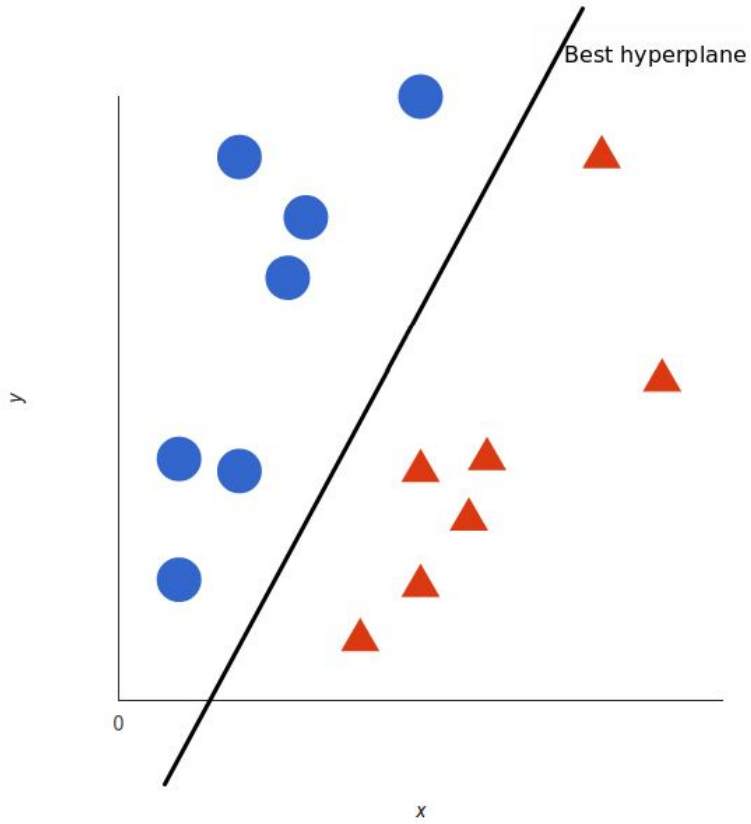
Dificultades

- Sobreajuste
- ⚖️ Desbalanceo
- Interpretación

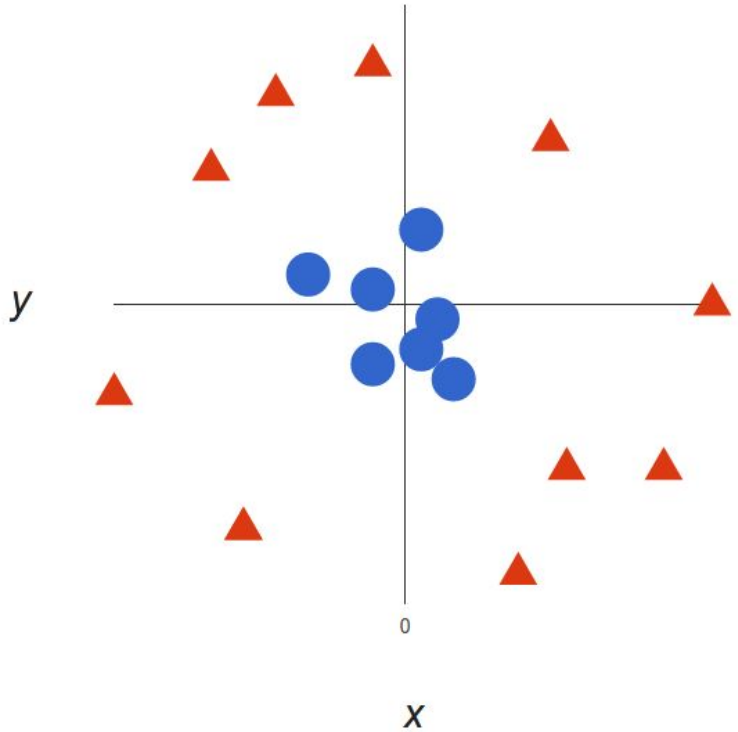
Clasificación

SVM

Lineal



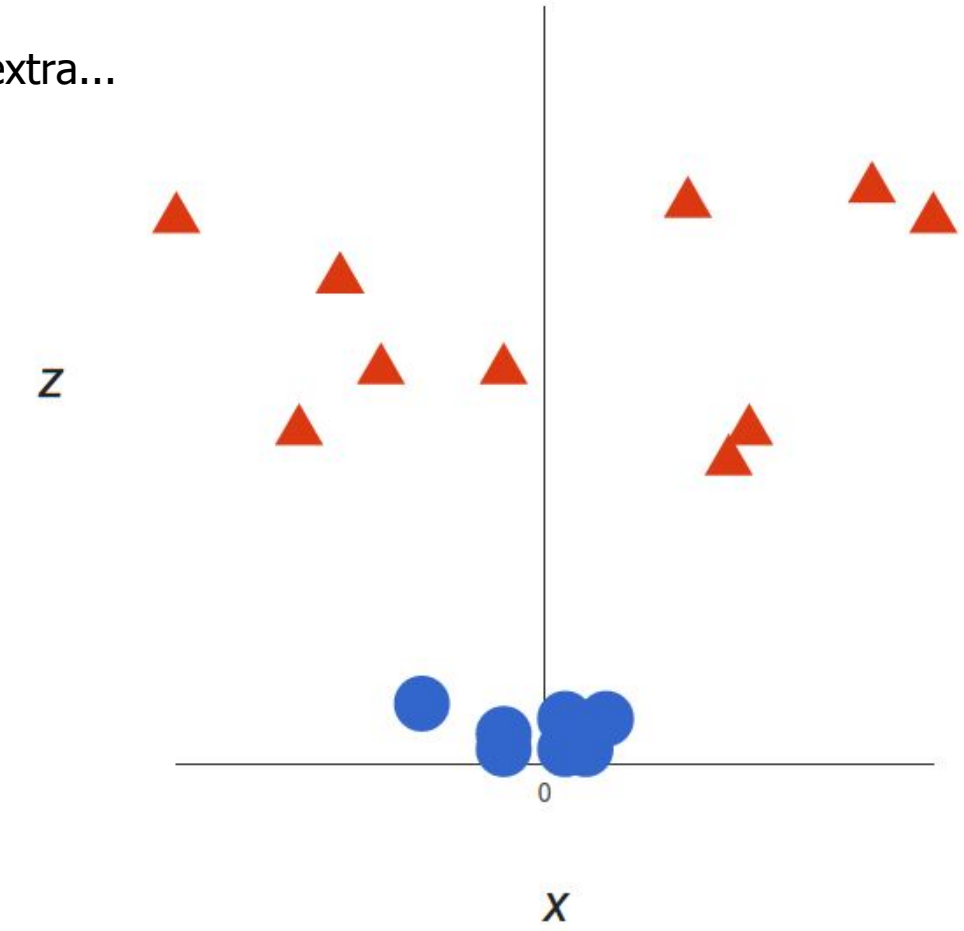
No lineal



Clasificación

SVM

Una dimensión extra...



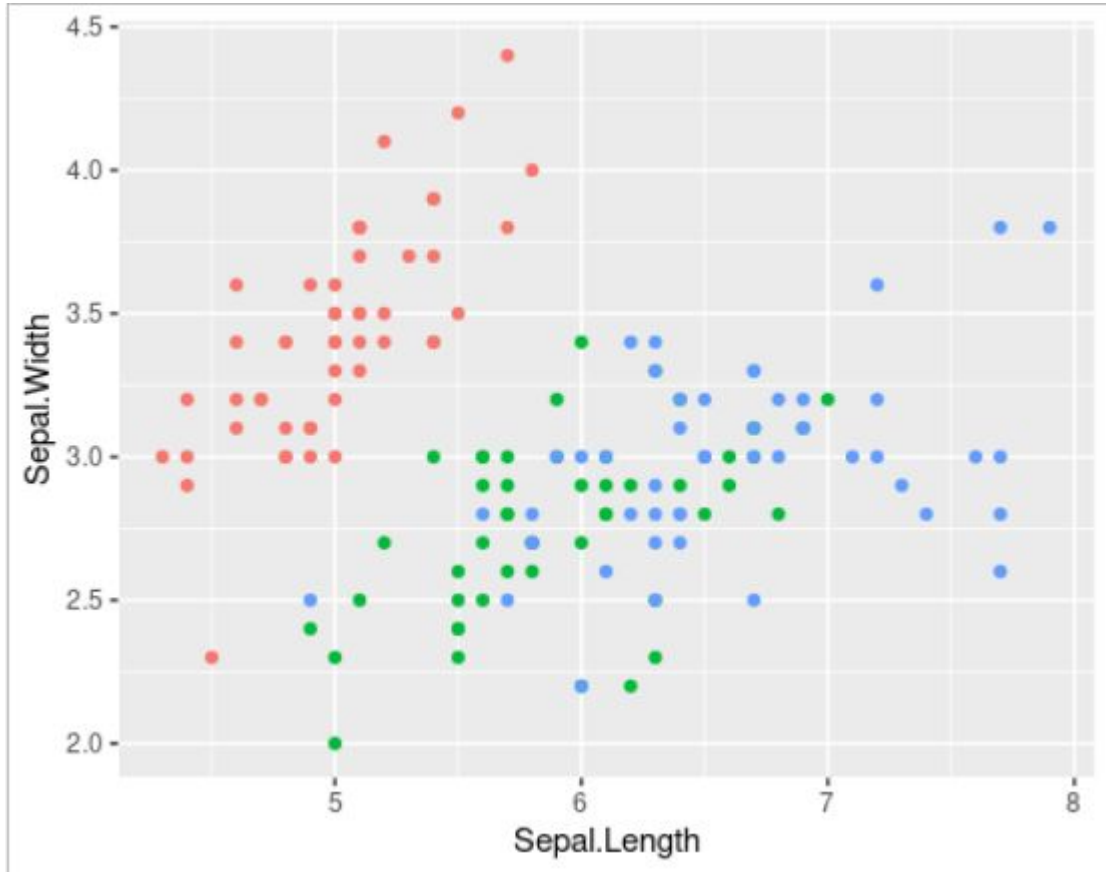
esden
Business School

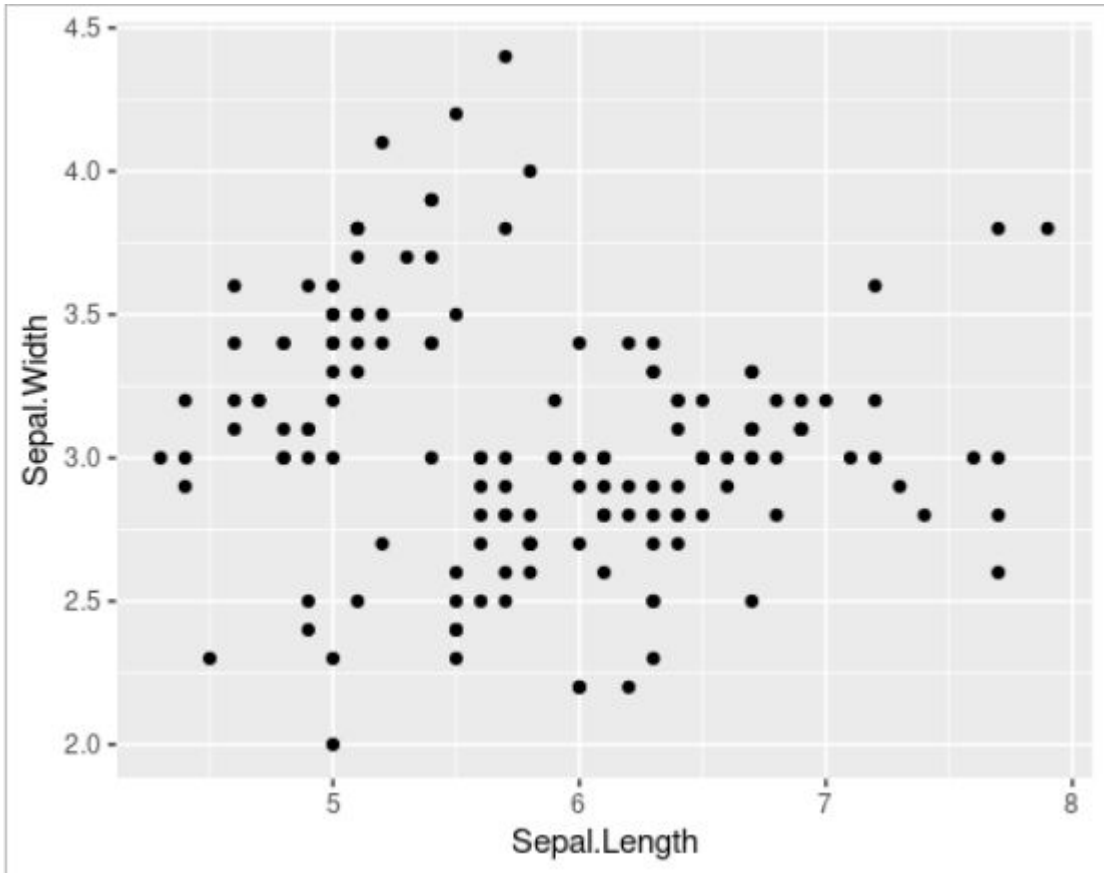
No supervisados

Cuando modelamos sin una referencia

No supervisados

Segmentación





No supervisados

k-means

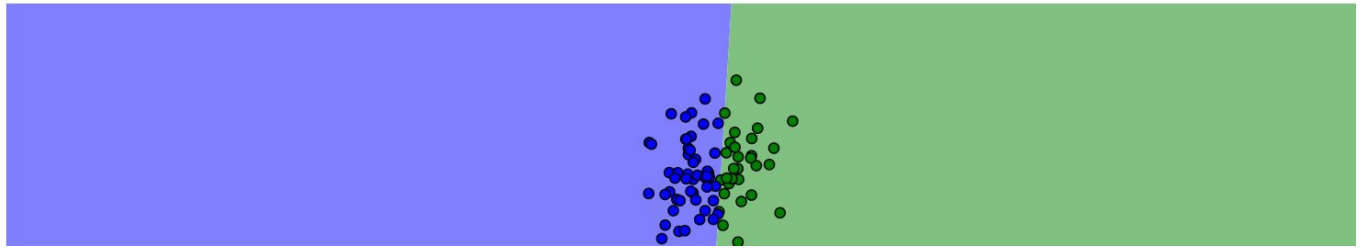
naftali harris

Blog | About | Contact | I'm Feeling Lucky

Visualizing K-Means Clustering

January 19, 2014

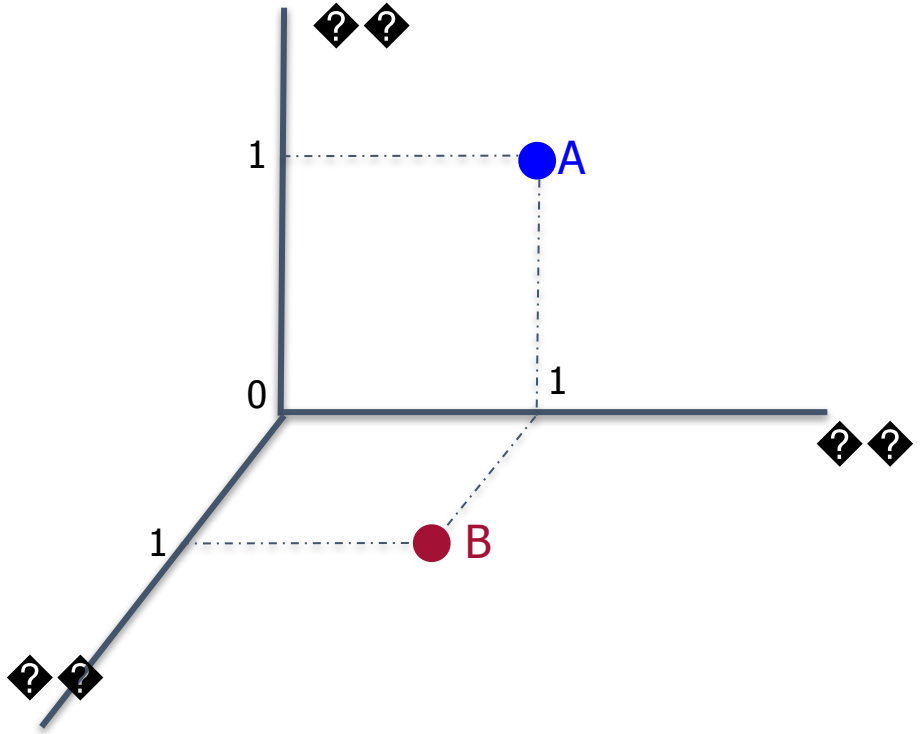
Suppose you plotted the screen width and height of all the devices accessing this website. You'd probably find that the points form three clumps: one clump with small dimensions, (smartphones), one with moderate dimensions, (tablets), and one with large dimensions, (laptops and desktops). Getting an algorithm to recognize these clumps of points without help is called *clustering*. To gain insight into how common clustering techniques work (and don't work), I've been making some visualizations that illustrate three fundamentally different approaches. This post, the first in this series of three, covers the k-means algorithm. To begin, click an initialization strategy below:



https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

No supervisados

Distancias



A = (1, 0, 1)
B = (1, 1, 0)

¿Cuál es la *distancia* entre ambos?

$$d_E = \sqrt{(1-1)^2 + (0-1)^2 + (1-0)^2} = 1,4142$$

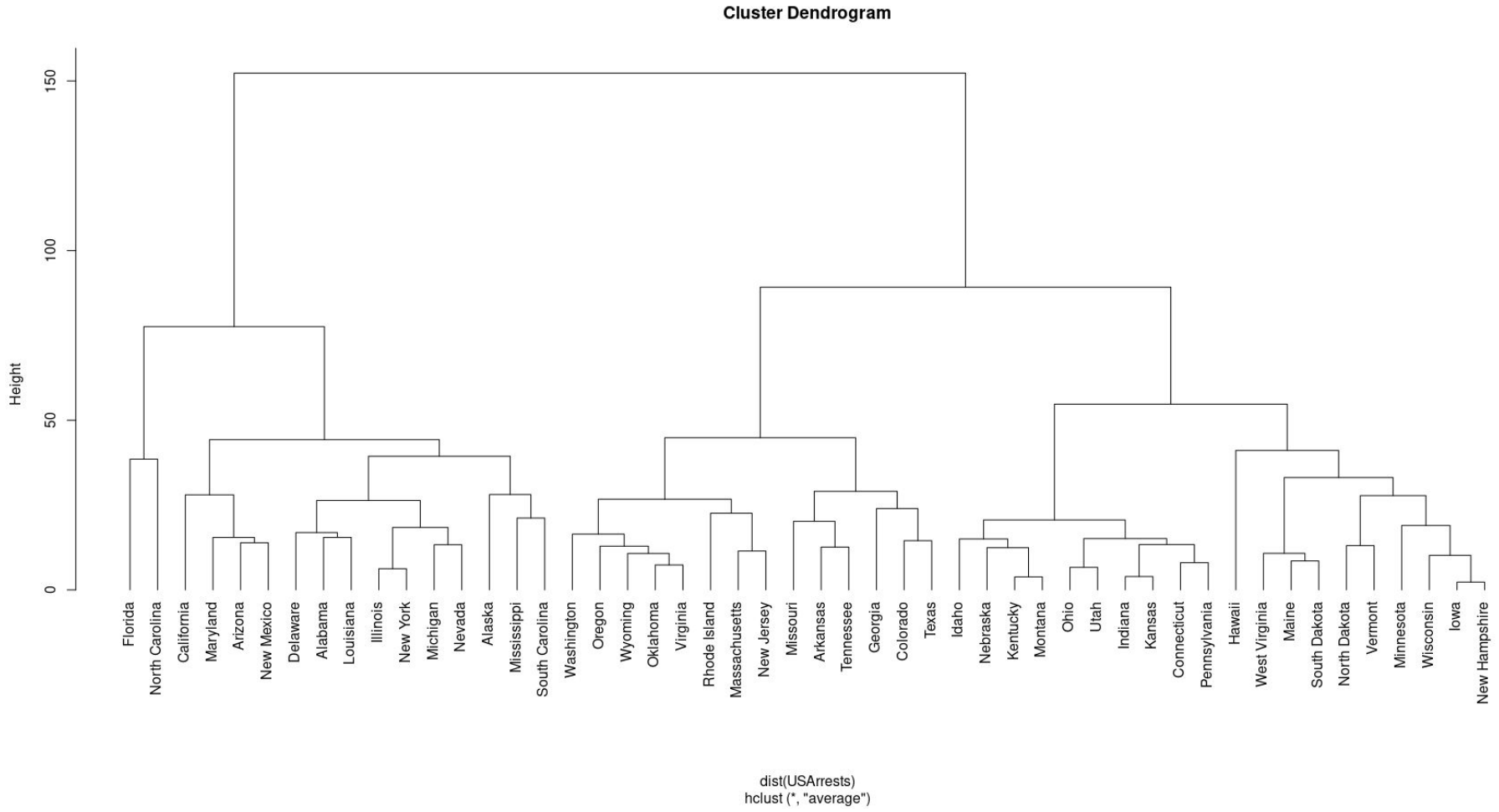
$$d_J = \frac{1}{1+1+1} = 0,3$$

Siguientes pasos

Cuando lo de hoy se nos queda corto

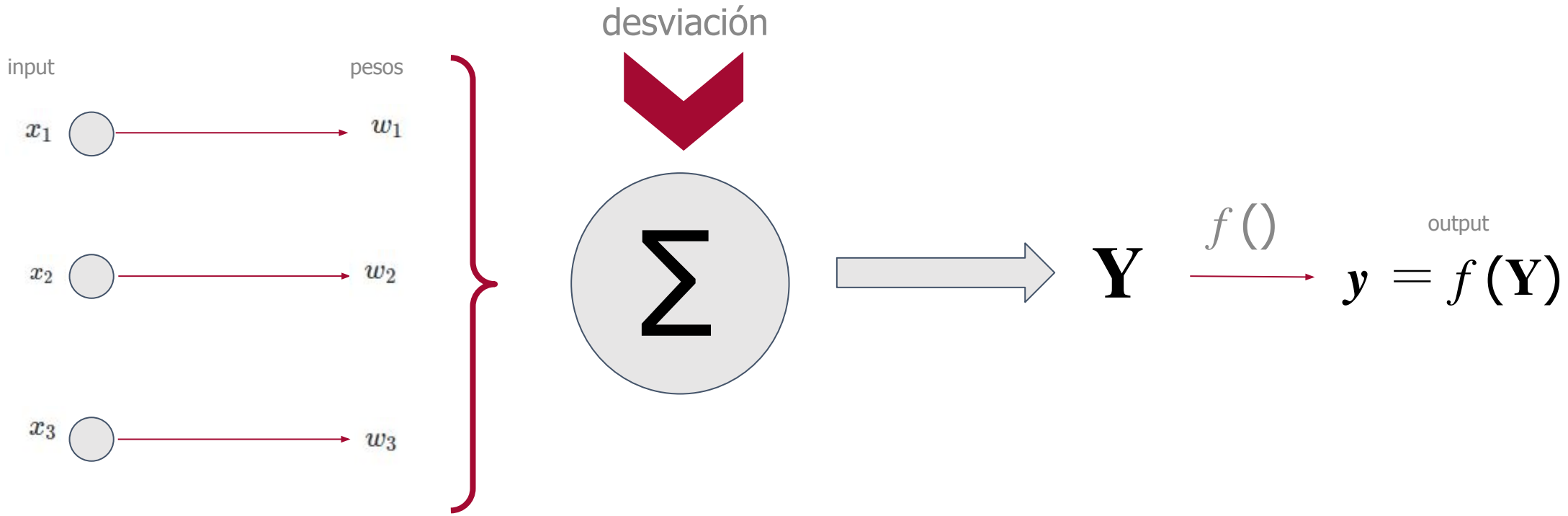
Siguientes pasos

Clusters jerárquicos



Siguientes pasos

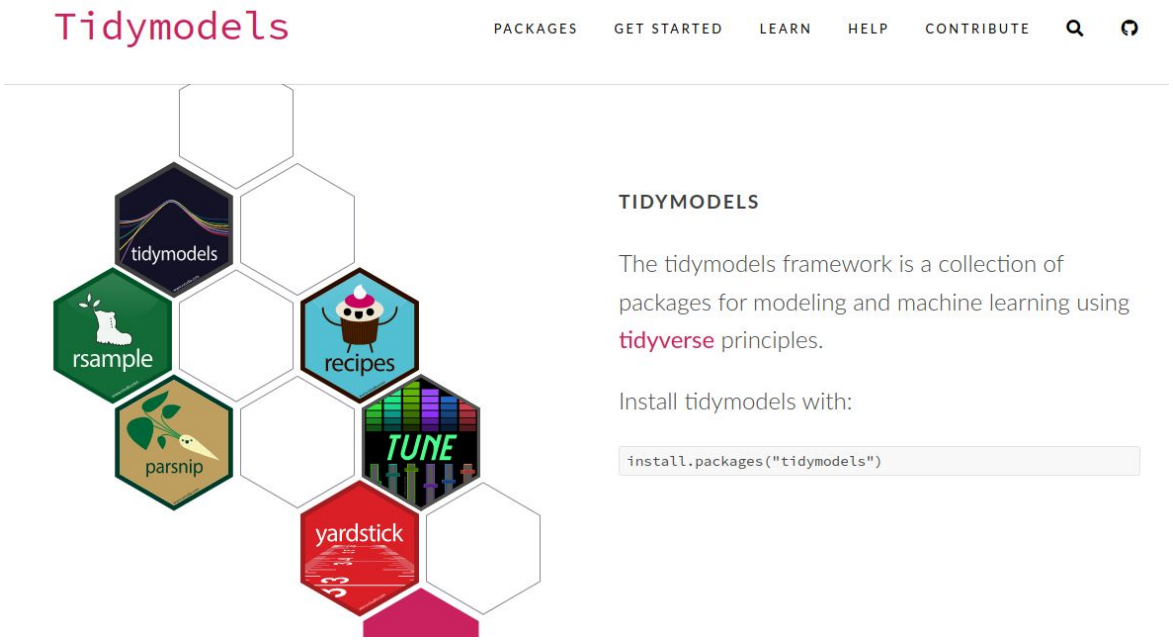
Redes neuronales



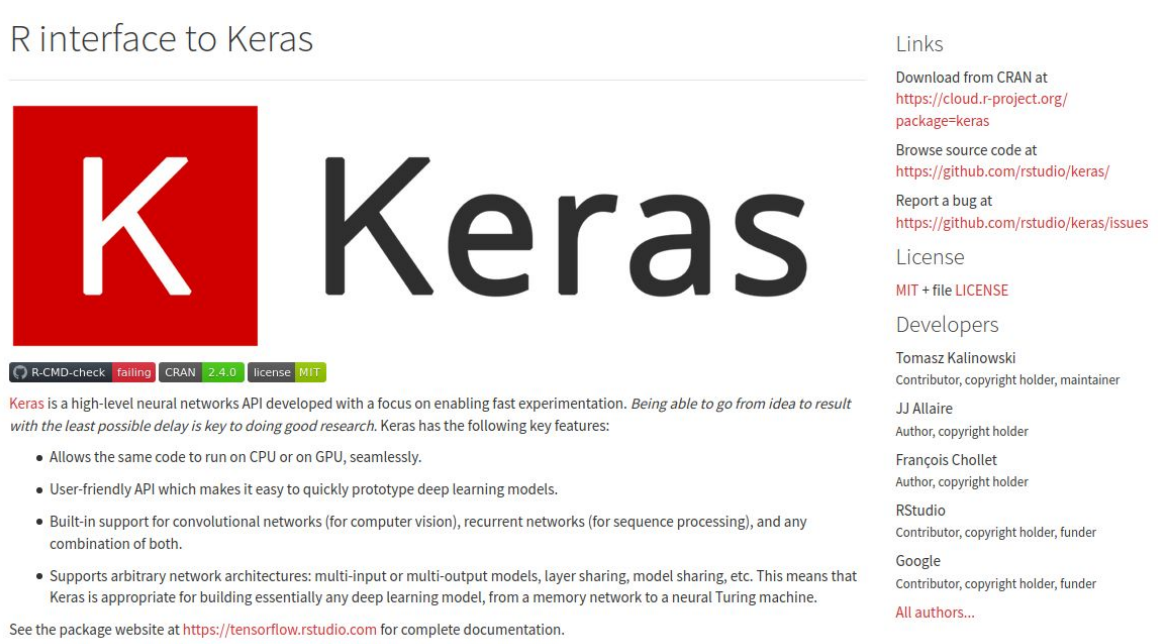
https://playground.tensorflow.org/

Siguientes pasos

Herramientas



https://www.tidymodels.org/



https://keras.rstudio.com/

Siguientes pasos

Influencers

- ★ Antonio Chinchón
- ★ Mariluz Congosto
- ★ José L. Cañadas
- ★ Danielle Navarro
- ★ T. L. Petersen
- ★ Joshua Kunst
- ★ Rosana Ferrero
- ★ Matt Dancho
- ★ Randy Olson
- ★ Rami Krispin
- ★ Pelayo Arbués
- ★ Hadley
- ★ Mara Averick
- ★ Julie Silge
- ★ Max Kuhn
- ★ Carlos G. Bellosta
- ★ Javier Á. Liébana
- ★ Kiko Llaneras

- ➢ RStudio
- ➢ R-Ladies
- ➢ Comunidad R Hispano
- ➢ Github
- ➢ Stackoverflow
- ➢ R4DS
- ➢ UMUR
- ➢ Datacamp

MADRID · BILBAO · BOGOTÁ · CDMX