



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. Siguiendo 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

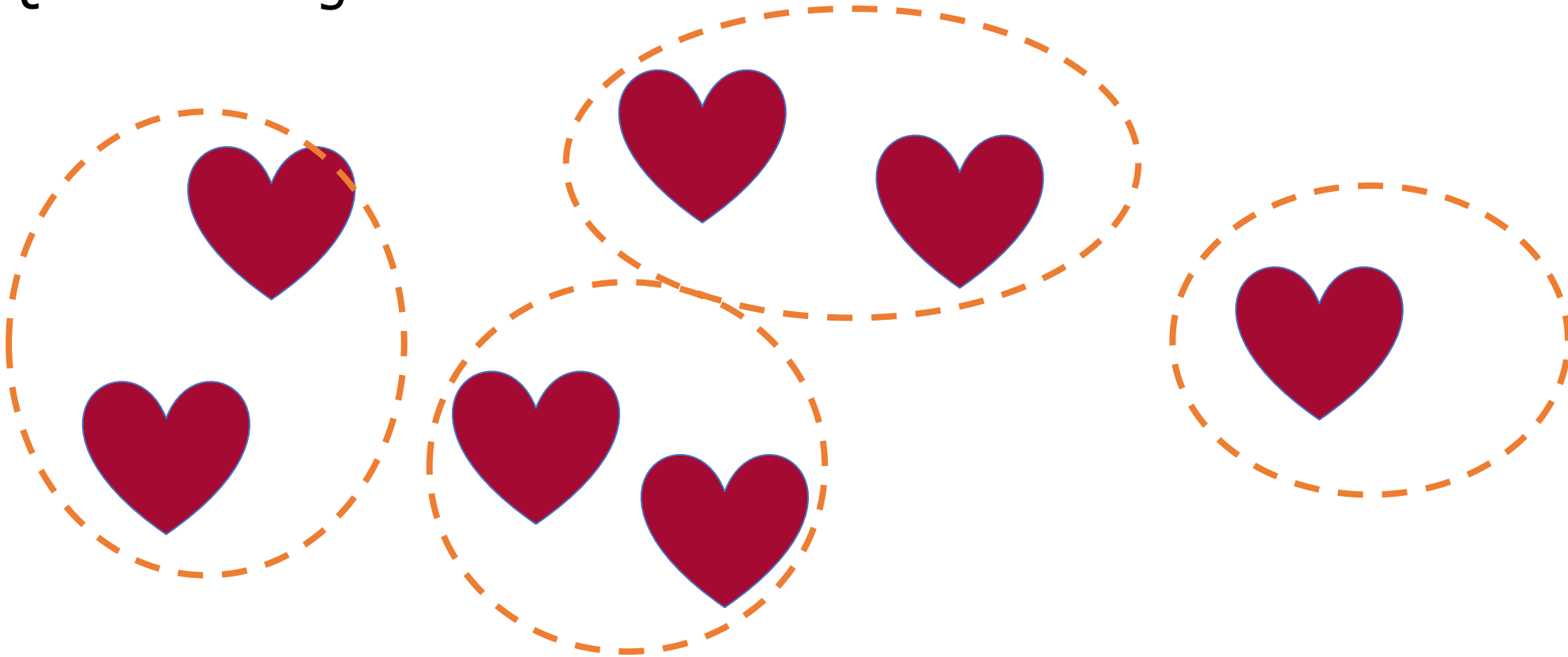
¿Qué es un modelo?



¿Qué es un modelo matemático?

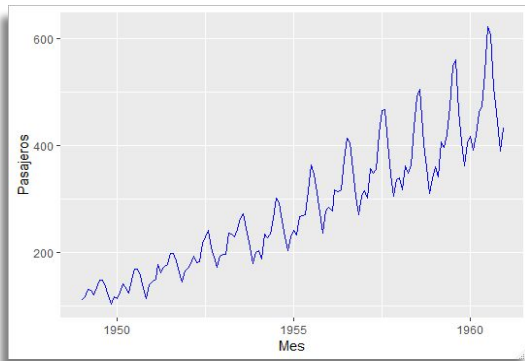
$$y = \alpha + \beta \cdot x$$

¿Qué es un algoritmo?

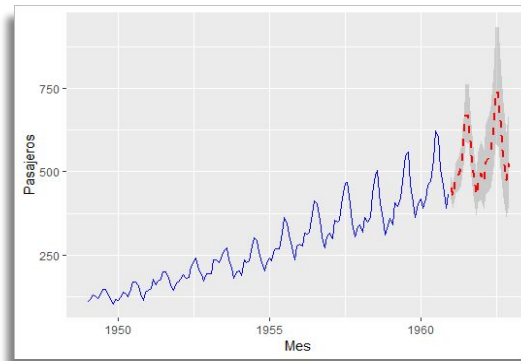


Sin histórico, no hay modelos

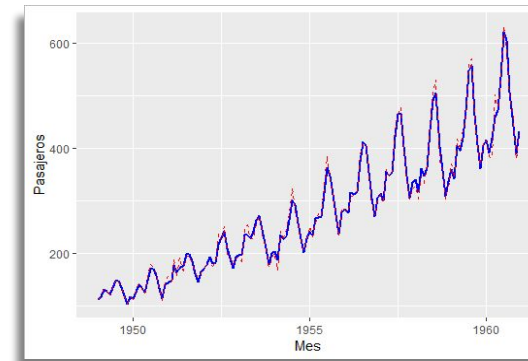
Paso 1. Datos históricos.



Paso 2. Ajustamos un modelo que aproxime el histórico.



Paso 3. Extrapolamos hacia el futuro.





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

[illegible][illegible]



Tools > Global options

Basic Graphics Advanced

R Sessions

Default working directory (when not in a project):
~ Browse...

☐ Restore most recently opened project at startup
☒ Restore previously open source documents at startup

Workspace

☐ Restore .RData into workspace at startup
Save workspace to .RData on exit: Never ▾

History

☐ Always save history (even when not saving .RData)
☐ Remove duplicate entries in history

Other

☐ Wrap around when navigating to previous/next tab
☒ Automatically notify me of updates to RStudio
☒ Send automated crash reports to RStudio

Regresión

Cuando modelamos datos continuos

Regresión

Regresión lineal

Coefficientes

P-valores

R2 y R2 ajustado

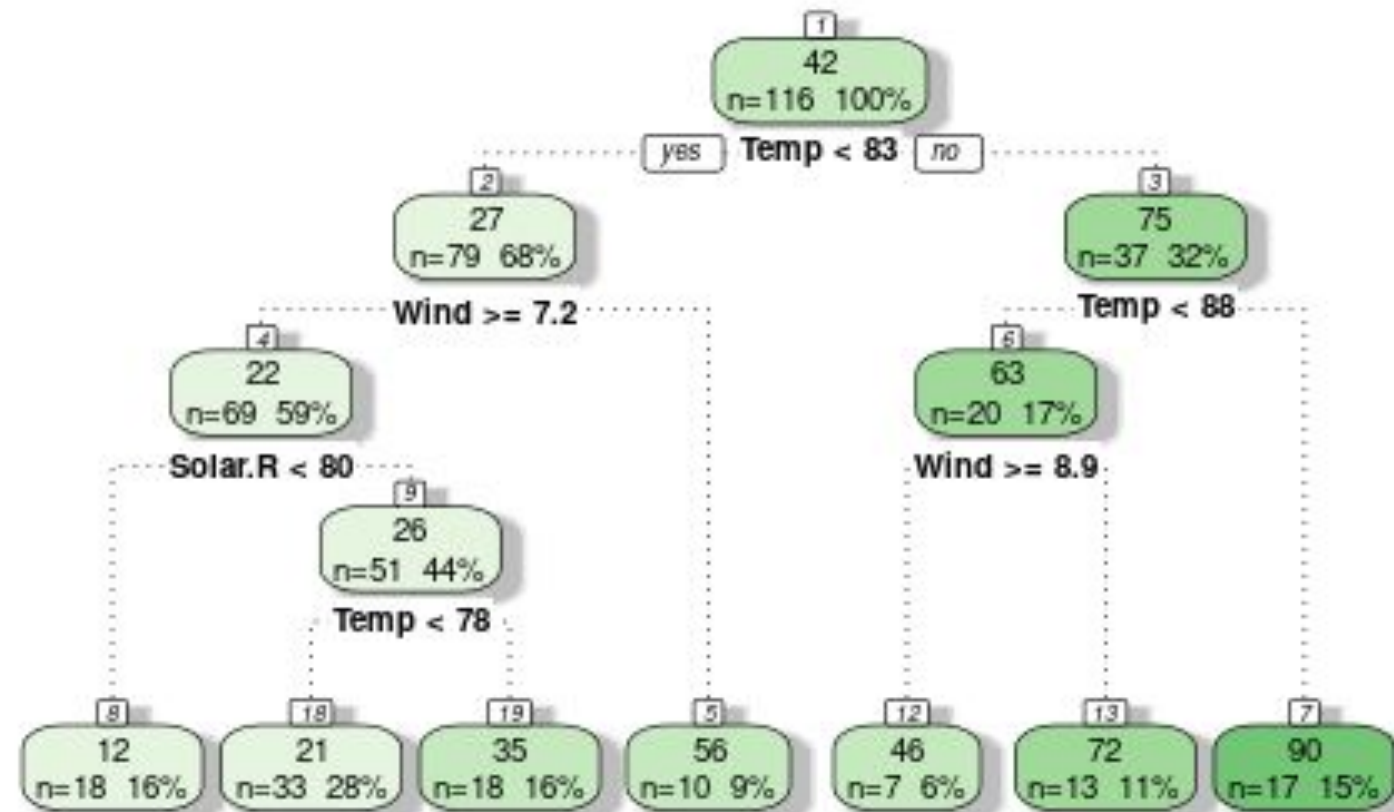
```
Call:
lm(formula = ventas ~ +1 + p2ola1_ad40 + p3_ola1_ad40 + p4_ola1_ad40 +
    p5_ola1_ad40 + p1_ola2_ad40 + p1_ola3_ad40 + dp + competencia1 +
    competencia2, data = df_consumo)

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

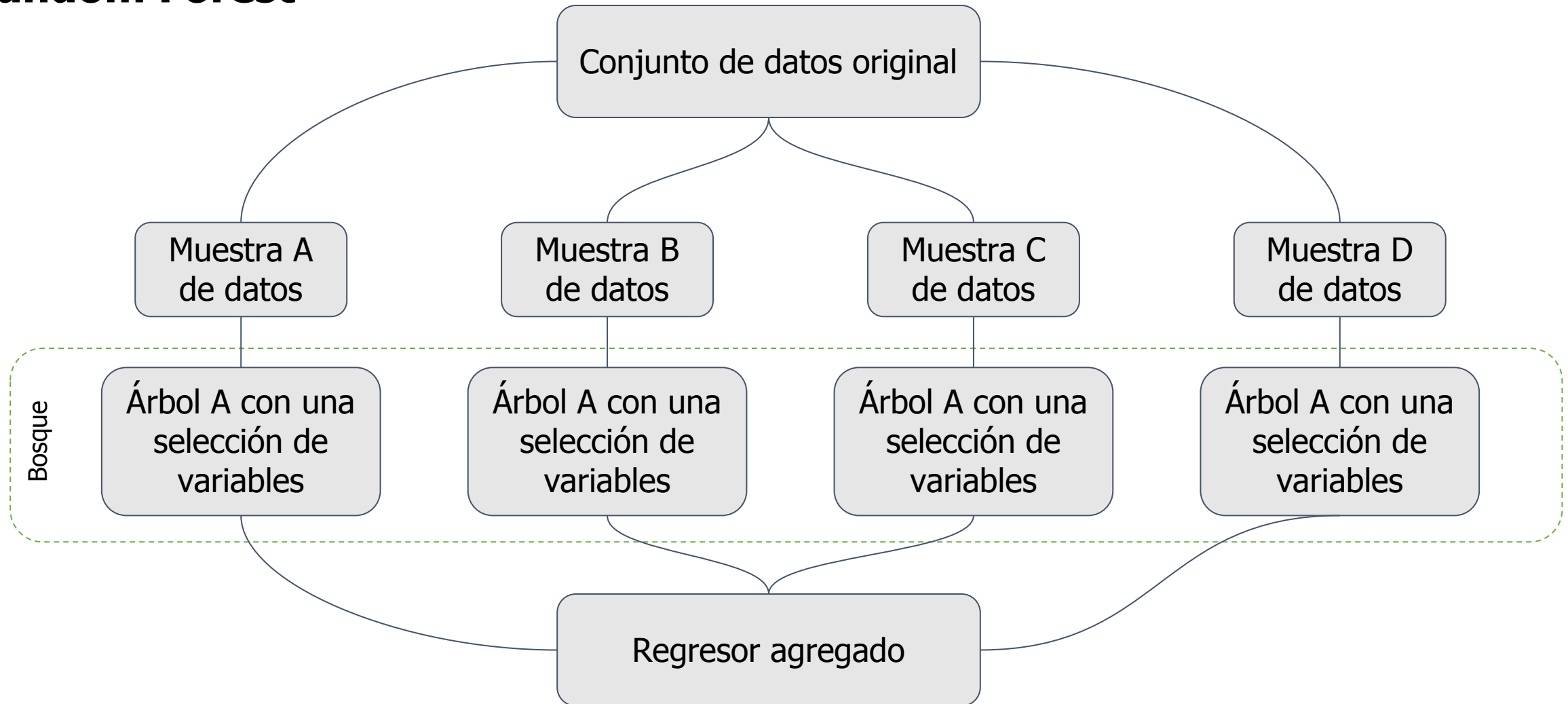
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.2851497   0.4408759  -11.988 < 2e-16 ***
p2ola1_ad40   0.0078508   0.0013079    6.003 1.43e-08 ***
p3_ola1_ad40   0.0039914   0.0005200    7.676 2.05e-12 ***
p4_ola1_ad40   0.0050948   0.0008755    5.819 3.53e-08 ***
p5_ola1_ad40   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
```

Árbol de regresión

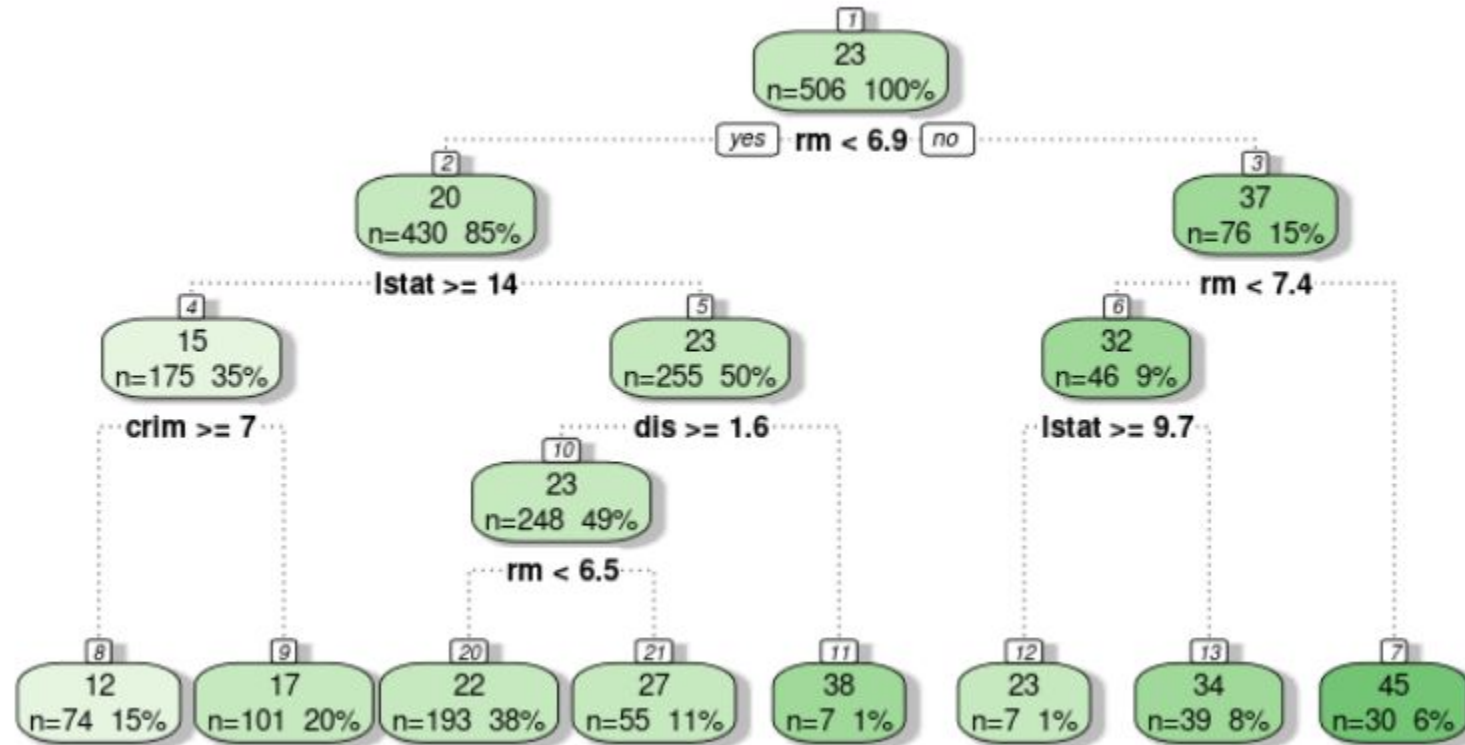


Random Forest



Regresión

Boston



Clasificación

Cuando modelamos categorías

Framingham

ten_year_chd.

padeció enfermedad 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 cigarrillos 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.

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

Dificultades

- Sobreajuste



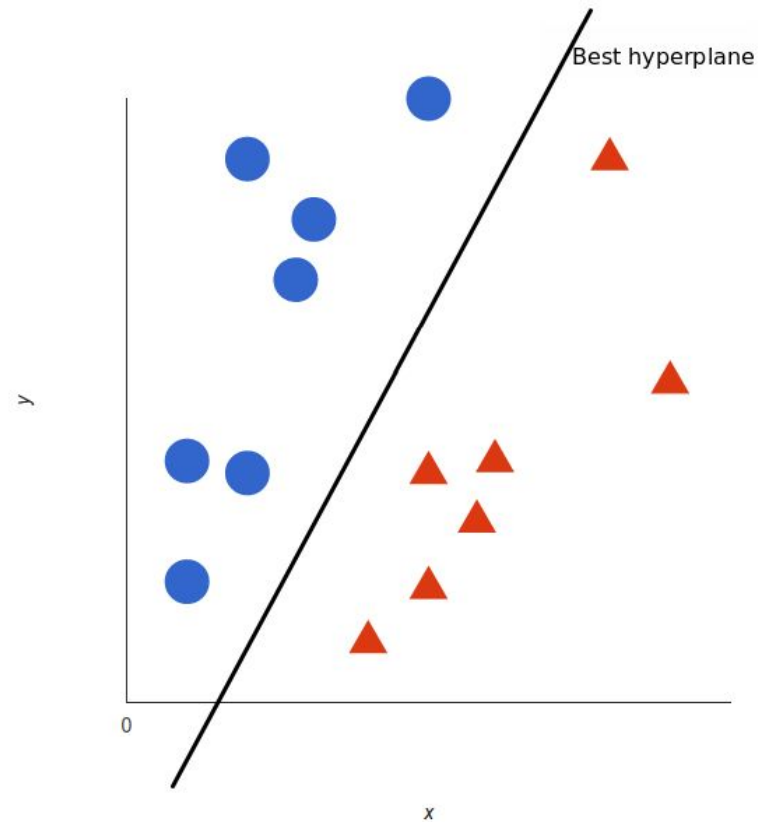
Desbalanceo

- Interpretación

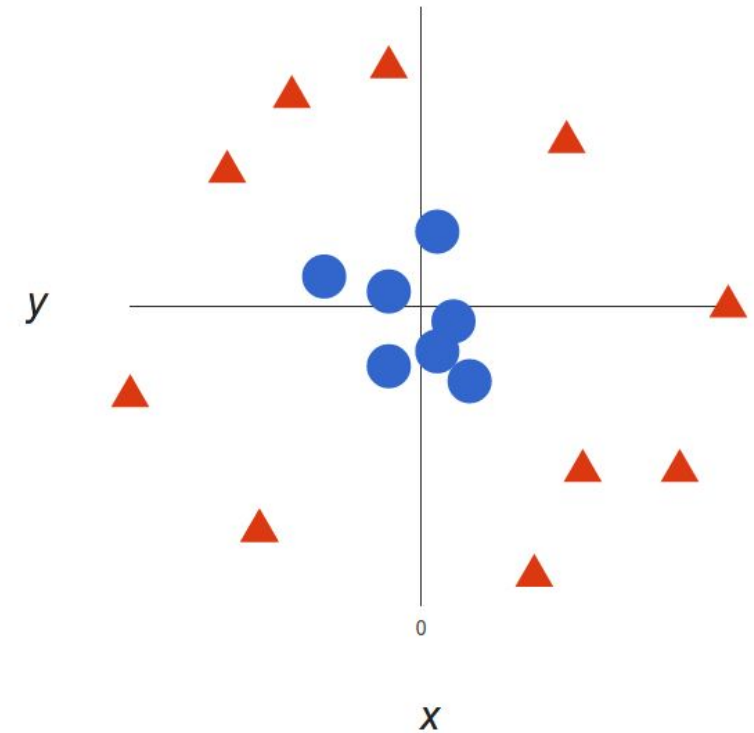
Clasificación

SVM

Lineal

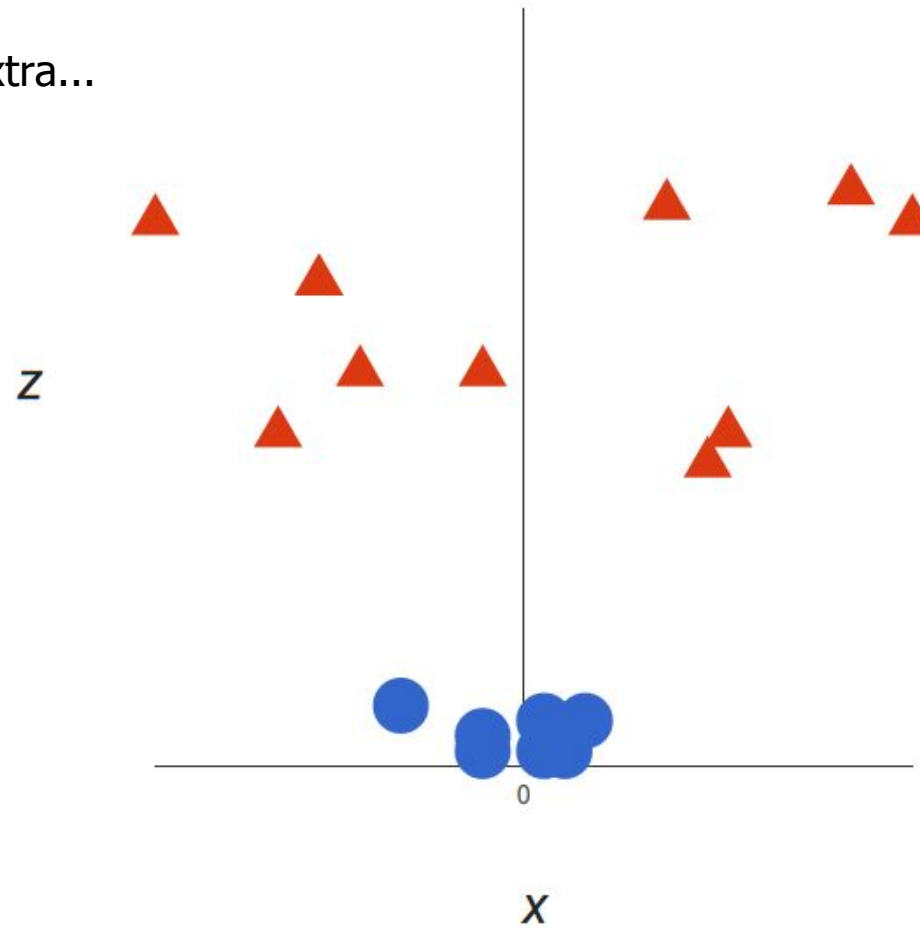


No lineal



SVM

Una dimensión extra...

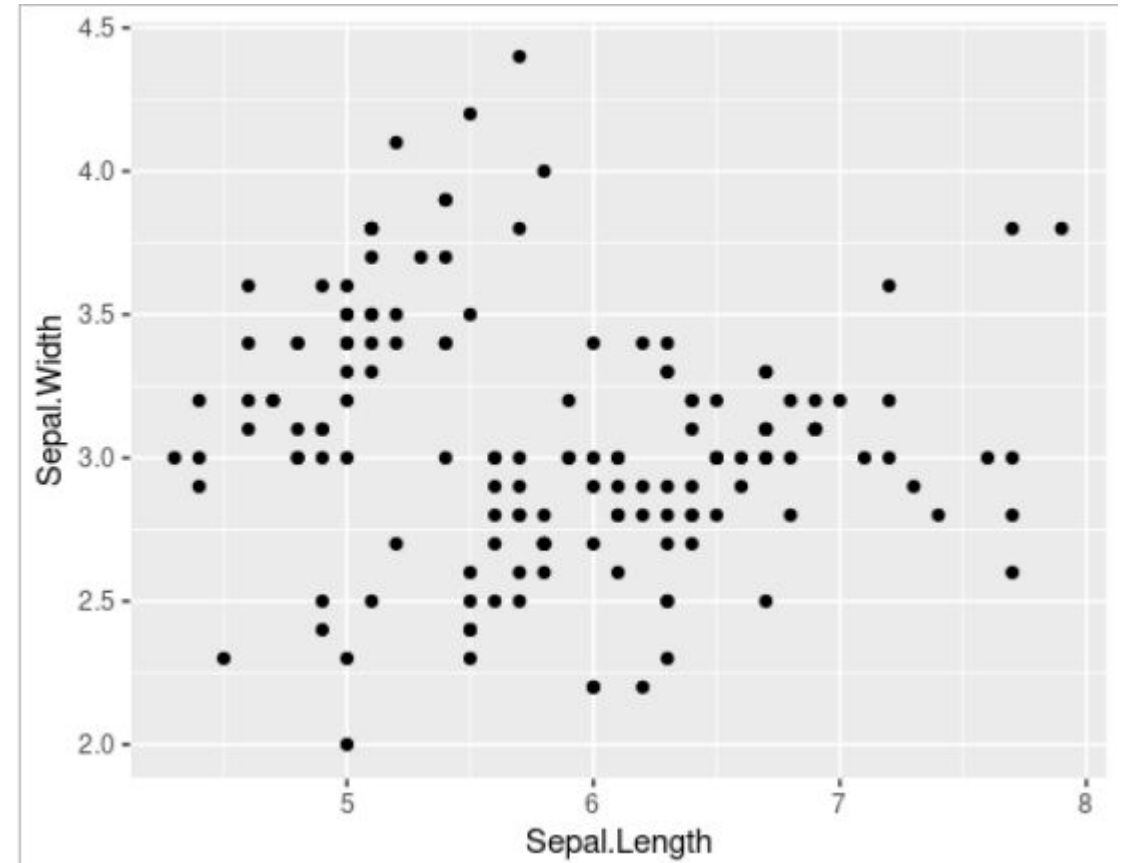
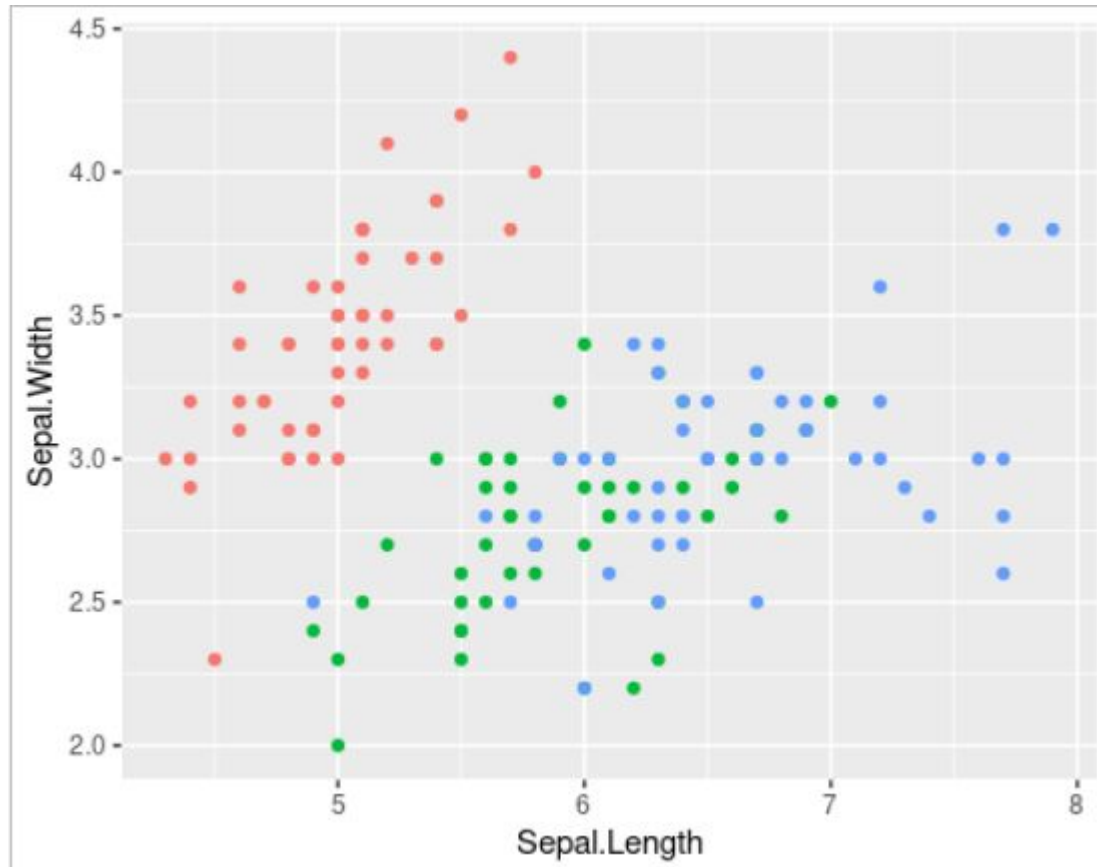


No supervisados

Cuando modelamos sin una referencia

No supervisados

Segmentación



No supervisados

k-means

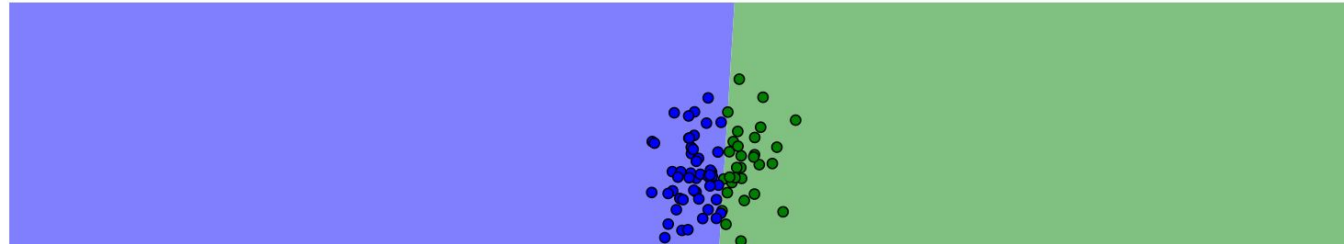
NH
naftali
harris

[Blog](#) [About](#) [Contact](#) [I'm Feeling Lucky](#)

Visualizing K-Means Clustering

January 19, 2014

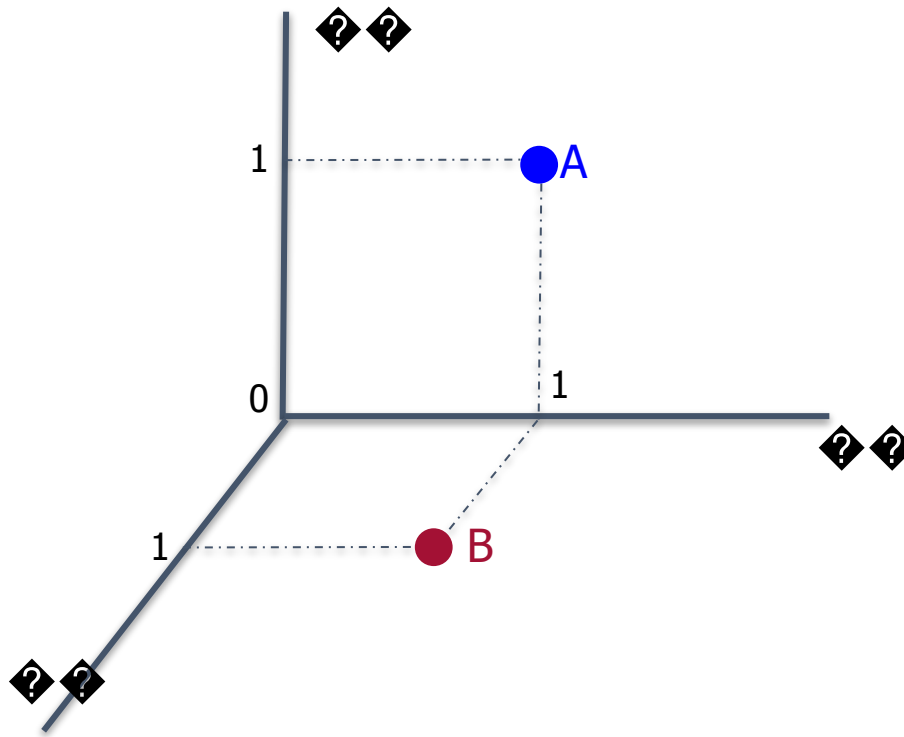
Suppose you plotted the screen width and height of all the devices accessing this website. You'd probably find that the points form three clumps: one clump with small dimensions, (smartphones), one with moderate dimensions, (tablets), and one with large dimensions, (laptops and desktops). Getting an 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



$A = (1, 0, 1)$ ¿Cuál es la *distancia*
 $B = (1, 1, 0)$ entre ambos?

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

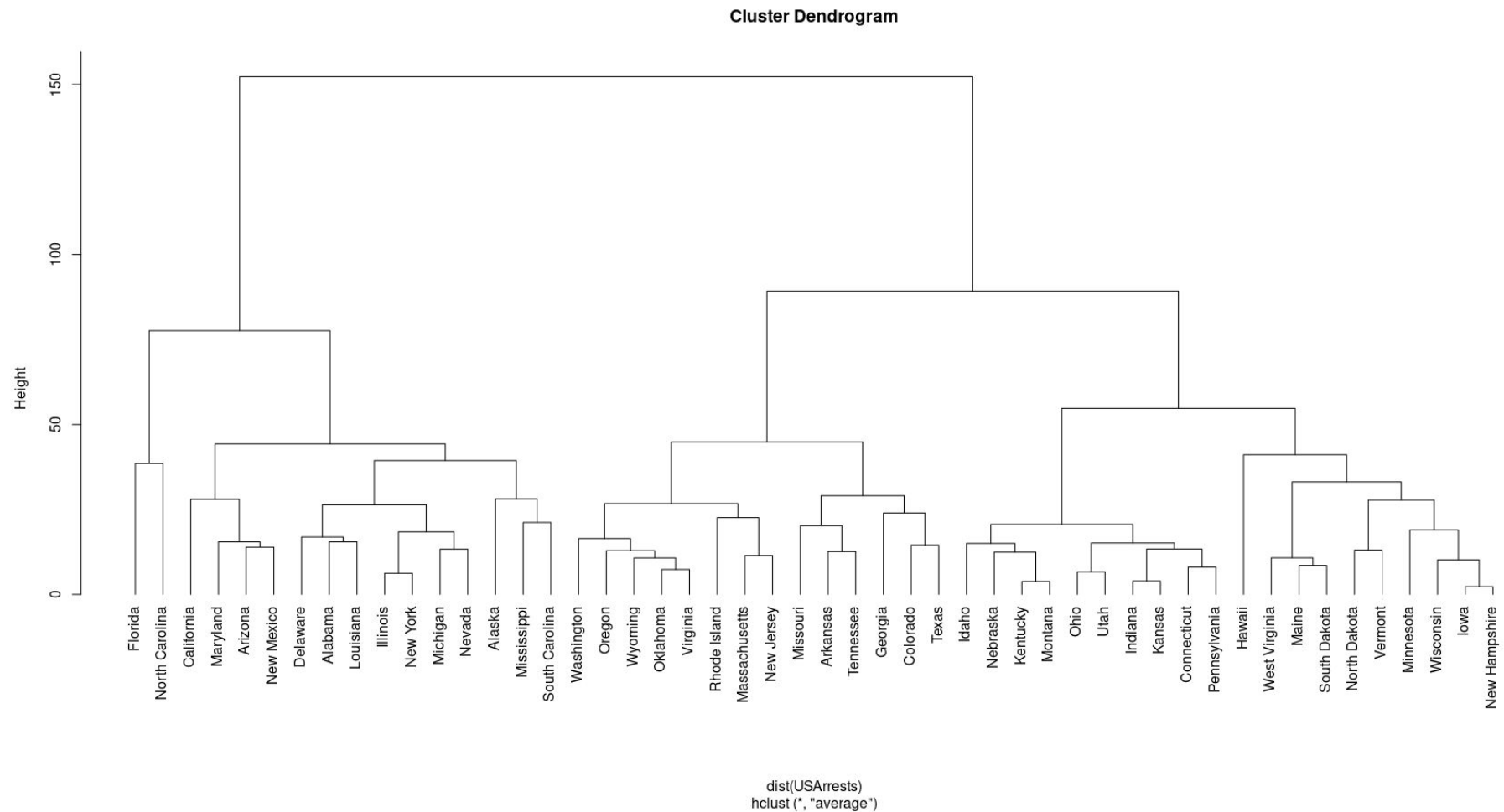
$$d_J = \frac{1}{1+1+1} = 0,3$$

Siguientes pasos

Cuando lo de hoy se nos queda corto

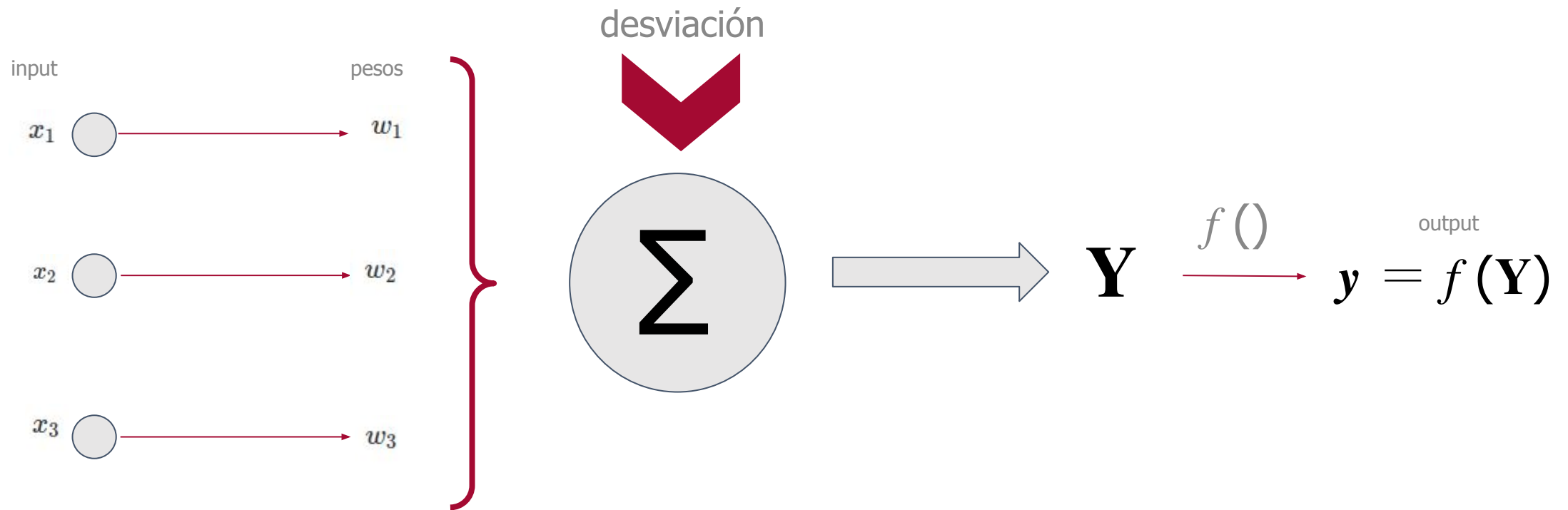
Siguientes pasos

Clusters jerárquicos



Siguientes pasos

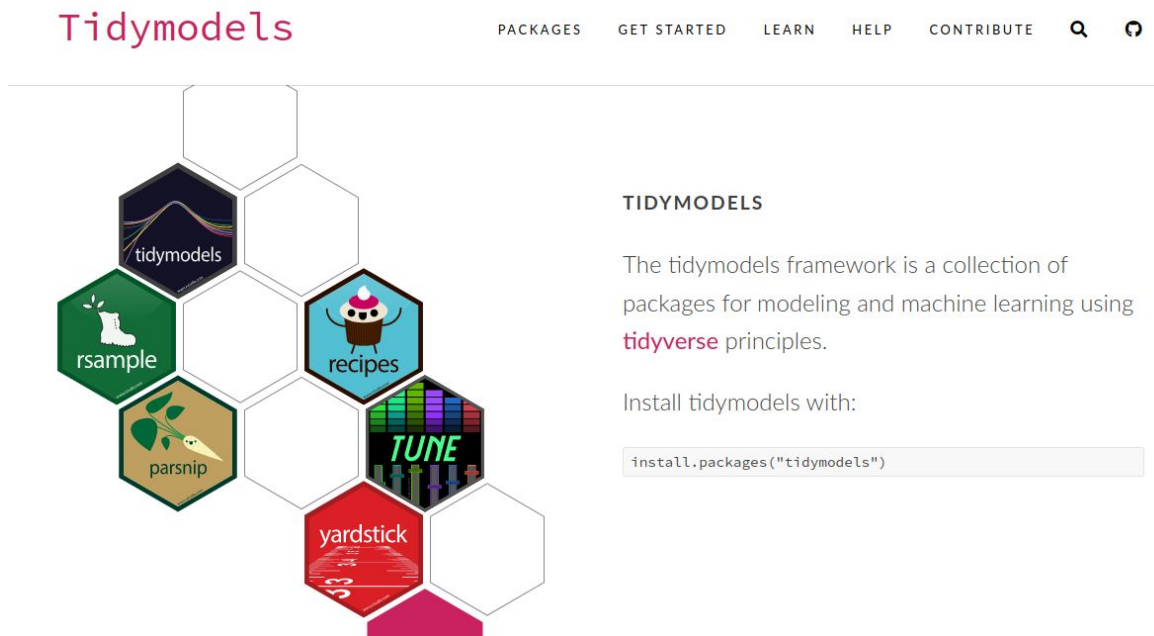
Redes neuronales



<https://playground.tensorflow.org/>

Siguientes pasos

Herramientas



<https://www.tidymodels.org/>

R interface to Keras



R-CMD-check failing CRAN 2.4.0 license MIT

Keras is a high-level neural networks API developed with a focus on enabling fast experimentation. *Being able to go from idea to result with the least possible delay is key to doing good research.* Keras has the following key features:

- Allows the same code to run on CPU or on GPU, seamlessly.
- User-friendly API which makes it easy to quickly prototype deep learning models.
- Built-in support for convolutional networks (for computer vision), recurrent networks (for sequence processing), and any combination of both.
- Supports arbitrary network architectures: multi-input or multi-output models, layer sharing, model sharing, etc. This means that Keras is appropriate for building essentially any deep learning model, from a memory network to a neural Turing machine.

See the package website at <https://tensorflow.rstudio.com> for complete documentation.

<https://keras.rstudio.com/>

Links

Download from CRAN at
<https://cloud.r-project.org/package=keras>

Browse source code at
<https://github.com/rstudio/keras/>

Report a bug at
<https://github.com/rstudio/keras/issues>

License

MIT + file LICENSE

Developers

Tomasz Kalinowski
Contributor, copyright holder, maintainer

JJ Allaire
Author, copyright holder

François Chollet
Author, copyright holder

RStudio
Contributor, copyright holder, funder

Google
Contributor, copyright holder, funder

All authors...

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