# CONCEPTOS BÁSICOS DE MACHINE LEARNING

Martín De la Fuente

# OBJETIVOS

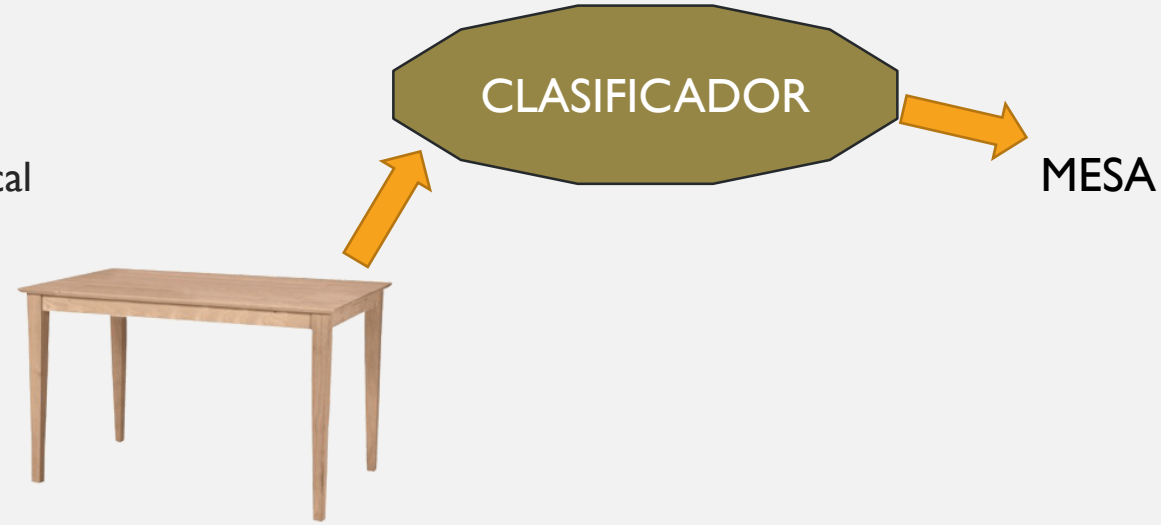
- Hacer resumen de los contenidos.
- Lograr entender bien qué estamos haciendo en la tarea.
- Manejar los términos que se usan en la tarea.
- Manejar conceptos para poder responder las preguntas de la tarea.
- Avanzar en su tarea y resolver dudas.

# CLASIFICACIÓN

# CLASIFICACIÓN

## ¿En qué consiste esta tarea?

- Sirve para detectar mediante un sistema automático la categoría (o clase) de un input.
- Por ejemplo:
  - Clasificar imágenes según su contenido
  - Clasificar canciones según su género musical
  - Clasificar dígitos escritos a mano
  - Clasificar objetos astronómicos según una observación de telescopio



# CLASIFICADORES

## ¿Qué clasificadores existen?

- Redes Neuronales (ANN)
- Vectores de Soporte (SVM)
- Regresión Logística (LR)
- Árboles de Decisión (DT)
- Vecinos Más Cercanos (KNN)
- Bayesiano Ingenuo (NB)

# CLASIFICADORES

## ¿Cómo construimos un clasificador?

- Debemos buscar la forma de "entrenar" a nuestro clasificador.
- Usamos datos "etiquetados" para que el clasificador aprenda a reconocer patrones (aprendizaje supervisado).

# CLASIFICADORES

MESA

SILLA

= ?

# DATOS

## ¿Cómo obtenemos los datos?

- Es una de las partes difíciles de esta tarea.
- En la mayoría de los casos son personas que tienen que hacer la clasificación de los datos.
- Por ejemplo:
  - Si queremos hacer un clasificador de calidad de vino, catadores de vinos
  - Si queremos hacer un clasificador de objetos astronómicos, astrónomos
  - Si queremos hacer un clasificador de imágenes, CAPTCHAs
  - Si queremos hacer un clasificador sobre los salarios que recibe cada persona, cuestionario

# DATOS

## ¿Cómo se estructuran los datos?

- Generalmente los tabulamos en tablas.
- Cada fila de la tabla es un dato, también llamada instancia o muestra del dataset.
- Cada columna de la tabla es una característica (*feature*) del dataset.
- Existe una columna especial que llamamos clase o categoría.
- Para tabular imágenes u otro tipo de datos podemos usar descriptores.

# Datos astronómicos

| objid | julian_date | ra | dec | redshift | u_filter | g_filter | r_filter | i_filter | z_filter | x_filter | camcol | field | plate | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6739133 | 53230 | 300.841762 | 76.5112824 | -0.00021952 | 19.19619 | 17.83329 | 17.52225 | 17.40237 | 17.35182 | 40131269 | 1 | 106 | 1660 | STAR |
| 4696263 | 53230 | 300.730508 | 76.5517315 | -8.36E-06 | 21.65541 | 19.13715 | 17.92577 | 17.44741 | 17.15818 | 98026415 | 1 | 106 | 1660 | STAR |
| 1905890 | 53230 | 300.871382 | 76.5305704 | 9.59E-05 | 20.70867 | 19.20954 | 18.55966 | 18.24395 | 18.10117 | 20605683 | 1 | 106 | 1660 | STAR |
| 2621494 | 53240 | 300.317409 | 76.374746 | -0.00024726 | 22.88806 | 21.209 | 19.9056 | 19.33555 | 19.08966 |  | 1 | 106 | 1661 | STAR |
| 3421446 | 53230 | 301.252332 | 76.3195196 | -0.0001315 | 17.82932 | 16.11081 | 15.39808 | 15.13612 | 15.00507 |  | 1 | 107 | 1660 | STAR |
| 5165444 | 53230 | 301.458518 | 76.4267657 | -0.00028757 | 14.23474 | 14.45194 | 14.75188 | 14.99473 | 15.24296 |  | 1 | 107 | 1660 | STAR |
| 8145190 | 53230 | 300.848872 | 76.3204622 | -0.0002666 | 21.69935 | 19.60817 | 18.18657 | 17.04788 | 16.44566 |  | 1 | 107 | 1660 | STAR |
| 2732193 | 53240 | 301.319371 | 76.3542733 | -0.00028874 | 25.33987 | 22.16937 | 20.70235 | 20.12803 | 19.72791 |  | 1 | 107 | 1661 | STAR |
| 3868589 | 53240 | 301.003481 | 76.4351085 | -0.00064419 | 21.77835 | 20.39711 | 19.72937 | 19.43841 | 19.2693 |  | 1 | 107 | 1661 | STAR |
| 6845117 | 53230 | 302.163901 | 76.388992 | -0.0002625 | 20.26735 | 18.5646 | 17.85914 | 17.58408 | 17.43464 |  | 1 | 108 | 1660 | STAR |
| 8465026 | 53230 | 302.174176 | 76.1933405 | -7.03E-05 | 21.54818 | 19.57721 | 18.43685 | 17.93544 | 17.62113 |  | 1 | 109 | 1660 | STAR |
| 2009504 | 53240 | 302.531797 | 76.3260686 | -2.79E-05 | 23.76005 | 21.57068 | 20.1001 | 19.45527 | 19.15964 |  | 1 | 109 | 1661 | STAR |
| 5734622 | 53230 | 302.53014 | 76.1376457 | -4.62E-05 | 21.58243 | 19.6114 | 18.55185 | 18.13491 | 17.91136 |  | 1 | 110 | 1660 | STAR |
| 8124788 | 53230 | 302.920973 | 76.3215062 | 2.68E-05 | 20.61019 | 18.13416 | 16.82791 | 16.22867 | 15.85798 |  | 1 | 110 | 1660 | STAR |
| 3893789 | 53240 | 303.625982 | 76.1874215 | -0.00020144 | 22.87485 | 21.75856 | 20.42418 | 19.75471 | 19.3883 |  | 1 | 111 | 1661 | STAR |
| 8931301 | 53230 | 303.76272 | 76.2041805 | -5.20E-05 | 17.27701 | 15.49217 | 14.6868 | 14.34704 | 14.15624 | 92214002 | 1 | 111 | 1660 | STAR |
| 4973464 | 53230 | 303.073864 | 76.1124891 | -0.00015582 | 21.94913 | 19.14192 | 17.67034 | 17.05692 | 16.69741 |  | 1 | 111 | 1660 | STAR |
| 7920184 | 53230 | 303.286843 | 76.0656598 | 6.70E-06 | 22.04301 | 19.78859 | 18.52149 | 18.03293 | 17.72046 |  | 1 | 111 | 1660 | STAR |
| 2164794 | 53240 | 303.506589 | 76.077405 | -0.00023373 | 22.4063 | 20.3364 | 19.52508 | 19.19709 | 19.02766 |  | 1 | 111 | 1661 | STAR |
| 5487131 | 53230 | 303.217422 | 76.1972257 | -9.25E-05 | 21.93067 | 20.09178 | 18.90102 | 18.33528 | 17.97431 |  | 1 | 111 | 1660 | STAR |
| 5927640 | 53240 | 303.354846 | 76.0528235 | -0.00043435 | 24.16451 | 22.21421 | 20.74198 | 20.24756 | 19.84714 |  | 1 | 111 | 1661 | STAR |
| 6297958 | 53230 | 303.852385 | 76.1738998 | -0.00023089 | 17.07861 | 15.71224 | 15.54605 | 15.49323 | 15.47547 |  | 1 | 112 | 1660 | STAR |
| 1728668 | 53240 | 304.020676 | 76.1601276 | -0.00054259 | 21.39911 | 19.83503 | 19.43652 | 19.24213 | 19.18858 |  | 1 | 112 | 1661 | STAR |
| 6343143 | 53230 | 303.983522 | 75.9810815 | -0.00014069 | 19.507 | 17.731 | 16.94764 | 16.65991 | 16.48766 |  | 1 | 112 | 1660 | STAR |
| 2612032 | 53240 | 303.98435 | 76.0978291 | -0.00014655 | 22.84097 | 20.48384 | 19.03249 | 18.43891 | 18.0931 |  | 1 | 112 | 1661 | STAR |
| 1405877 | 53240 | 303.941362 | 76.0801423 | -0.00029587 | 23.18969 | 20.9254 | 19.63644 | 19.03506 | 18.61127 | 20727395 | 1 | 112 | 1661 | STAR |
| 5771574 | 53240 | 304.152653 | 76.0944185 | -0.00019201 | 23.9366 | 21.53848 | 20.07633 | 19.51465 | 19.11936 | 14140728 | 1 | 112 | 1661 | STAR |
| 1410471 | 53230 | 304.531555 | 76.1320652 | -0.00022718 | 19.56016 | 18.1636 | 17.49738 | 17.2079 | 17.036 | 30894670 | 1 | 113 | 1660 | STAR |
| 3804784 | 53230 | 304.373259 | 75.9887762 | -0.00028019 | 21.98413 | 19.17101 | 17.96169 | 17.46724 | 17.22819 |  | 1 | 113 | 1660 | STAR |
| 8551198 | 53230 | 304.602721 | 76.119609 | -0.00016839 | 21.64969 | 19.46481 | 18.30491 | 17.83545 | 17.54242 |  | 1 | 113 | 1660 | STAR |
| 3498800 | 53240 | 304.458928 | 75.9339433 | -0.00031972 | 16.92797 | 15.63552 | 15.10081 | 14.89095 | 14.79415 |  | 1 | 113 | 1661 | STAR |
| 9187177 | 53230 | 304.447666 | 76.0485287 | -1.22E-05 | 19.25177 | 16.60164 | 15.20536 | 14.54854 | 14.19345 | 92679336 | 1 | 113 | 1660 | STAR |
| 2518609 | 53230 | 304.535053 | 75.9621095 | -0.0003306 | 20.77139 | 19.33941 | 18.614 | 18.37783 | 18.19239 | 90707782 | 1 | 113 | 1660 | STAR |

# Datos astronómicos: identificar la clase

| objid | julian_date | ra | dec | redshift | u_filter | g_filter | r_filter | i_filter | z_filter | x_filter | camcol | field | plate | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6739133 | 53230 | 300.841762 | 76.5112824 | -0.00021952 | 19.19619 | 17.83329 | 17.52225 | 17.40237 | 17.35182 | 40131269 | 1 | 106 | 1660 | STAR |
| 4696263 | 53230 | 300.730508 | 76.5517315 | -8.36E-06 | 21.65541 | 19.13715 | 17.92577 | 17.44741 | 17.15818 | 98026415 | 1 | 106 | 1660 | STAR |
| 1905890 | 53230 | 300.871382 | 76.5305704 | 9.59E-05 | 20.70867 | 19.20954 | 18.55966 | 18.24395 | 18.10117 | 20605683 | 1 | 106 | 1660 | STAR |
| 2621494 | 53240 | 300.317409 | 76.374746 | -0.00024726 | 22.88806 | 21.209 | 19.9056 | 19.33555 | 19.08966 | | 1 | 106 | 1661 | STAR |
| 3421446 | 53230 | 301.252332 | 76.3195196 | -0.0001315 | 17.82932 | 16.11081 | 15.39808 | 15.13612 | 15.00507 | | 1 | 107 | 1660 | STAR |
| 5165444 | 53230 | 301.458518 | 76.4267657 | -0.00028757 | 14.23474 | 14.45194 | 14.75188 | 14.99473 | 15.24296 | | 1 | 107 | 1660 | STAR |
| 8145190 | 53230 | 300.848872 | 76.3204622 | -0.0002666 | 21.69935 | 19.60817 | 18.18657 | 17.04788 | 16.44566 | | 1 | 107 | 1660 | STAR |
| 2732193 | 53240 | 301.319371 | 76.3542733 | -0.00028874 | 25.33987 | 22.16937 | 20.70235 | 20.12803 | 19.72791 | | 1 | 107 | 1661 | STAR |
| 3868589 | 53240 | 301.003481 | 76.4351085 | -0.00064419 | 21.77835 | 20.39711 | 19.72937 | 19.43841 | 19.2693 | | 1 | 107 | 1661 | STAR |
| 6845117 | 53230 | 302.163901 | 76.388992 | -0.0002625 | 20.26735 | 18.5646 | 17.85914 | 17.58408 | 17.43464 | | 1 | 108 | 1660 | STAR |
| 8465026 | 53230 | 302.174176 | 76.1933405 | -7.03E-05 | 21.54818 | 19.57721 | 18.43685 | 17.93544 | 17.62113 | | 1 | 109 | 1660 | STAR |
| 2009504 | 53240 | 302.531797 | 76.3260686 | -2.79E-05 | 23.76005 | 21.57068 | 20.1001 | 19.45527 | 19.15964 | | 1 | 109 | 1661 | STAR |
| 5734622 | 53230 | 302.53014 | 76.1376457 | -4.62E-05 | 21.58243 | 19.6114 | 18.55185 | 18.13491 | 17.91136 | | 1 | 110 | 1660 | STAR |
| 8124788 | 53230 | 302.920973 | 76.3215062 | 2.68E-05 | 20.61019 | 18.13416 | 16.82791 | 16.22867 | 15.85798 | | 1 | 110 | 1660 | STAR |
| 3893789 | 53240 | 303.625982 | 76.1874215 | -0.00020144 | 22.87485 | 21.75856 | 20.42418 | 19.75471 | 19.3883 | | 1 | 111 | 1661 | STAR |
| 8931301 | 53230 | 303.76272 | 76.2041805 | -5.20E-05 | 17.27701 | 15.49217 | 14.6868 | 14.34704 | 14.15624 | 92214002 | 1 | 111 | 1660 | STAR |
| 4973464 | 53230 | 303.073864 | 76.1124891 | -0.00015582 | 21.94913 | 19.14192 | 17.67034 | 17.05692 | 16.69741 | | 1 | 111 | 1660 | STAR |
| 7920184 | 53230 | 303.286843 | 76.0656598 | 6.70E-06 | 22.04301 | 19.78859 | 18.52149 | 18.03293 | 17.72046 | | 1 | 111 | 1660 | STAR |
| 2164794 | 53240 | 303.506589 | 76.077405 | -0.00023373 | 22.4063 | 20.3364 | 19.52508 | 19.19709 | 19.02766 | | 1 | 111 | 1661 | STAR |
| 5487131 | 53230 | 303.217422 | 76.1972257 | -9.25E-05 | 21.93067 | 20.09178 | 18.90102 | 18.33528 | 17.97431 | | 1 | 111 | 1660 | STAR |
| 5927640 | 53240 | 303.354846 | 76.0528235 | -0.00043435 | 24.16451 | 22.21421 | 20.74198 | 20.24756 | 19.84714 | | 1 | 111 | 1661 | STAR |
| 6297958 | 53230 | 303.852385 | 76.1738998 | -0.00023089 | 17.07861 | 15.71224 | 15.54605 | 15.49323 | 15.47547 | | 1 | 112 | 1660 | STAR |
| 1728668 | 53240 | 304.020676 | 76.1601276 | -0.00054259 | 21.39911 | 19.83503 | 19.43652 | 19.24213 | 19.18858 | | 1 | 112 | 1661 | STAR |
| 6343143 | 53230 | 303.983522 | 75.9810815 | -0.00014069 | 19.507 | 17.731 | 16.94764 | 16.65991 | 16.48766 | | 1 | 112 | 1660 | STAR |
| 2612032 | 53240 | 303.98435 | 76.0978291 | -0.00014655 | 22.84097 | 20.48384 | 19.03249 | 18.43891 | 18.0931 | | 1 | 112 | 1661 | STAR |
| 1405877 | 53240 | 303.941362 | 76.0801423 | -0.00029587 | 23.18969 | 20.9254 | 19.63644 | 19.03506 | 18.61127 | 20727395 | 1 | 112 | 1661 | STAR |
| 5771574 | 53240 | 304.152653 | 76.0944185 | -0.00019201 | 23.9366 | 21.53848 | 20.07633 | 19.51465 | 19.11936 | 14140728 | 1 | 112 | 1661 | STAR |
| 1410471 | 53230 | 304.531555 | 76.1320652 | -0.00022718 | 19.56016 | 18.1636 | 17.49738 | 17.2079 | 17.036 | 30894670 | 1 | 113 | 1660 | STAR |
| 3804784 | 53230 | 304.373259 | 75.9887762 | -0.00028019 | 21.98413 | 19.17101 | 17.96169 | 17.46724 | 17.22819 | | 1 | 113 | 1660 | STAR |
| 8551198 | 53230 | 304.602721 | 76.119609 | -0.00016839 | 21.64969 | 19.46481 | 18.30491 | 17.83545 | 17.54242 | | 1 | 113 | 1660 | STAR |
| 3498800 | 53240 | 304.458928 | 75.9339433 | -0.00031972 | 16.92797 | 15.63552 | 15.10081 | 14.89095 | 14.79415 | | 1 | 113 | 1661 | STAR |
| 9187177 | 53230 | 304.447666 | 76.0485287 | -1.22E-05 | 19.25177 | 16.60164 | 15.20536 | 14.54854 | 14.19345 | 92679336 | 1 | 113 | 1660 | STAR |
| 2518609 | 53230 | 304.535053 | 75.9621095 | -0.0003306 | 20.77139 | 19.33941 | 18.614 | 18.37783 | 18.19239 | 90707782 | 1 | 113 | 1660 | STAR |

Características

Clase

# DATOS

## ¿Siempre usamos todas las características?

- No siempre, a veces queremos seleccionar la mejores.
- Podemos hacer una selección a priori, quitando datos que definitivamente no aportan (identificadores, fechas de cuando se obtuvo la información, etc)
- Podemos hacer una selección automática, con algoritmos que verifican la correlación entre cada característica y la clase. A mayor correlación, más valiosa la característica.

# Datos astronómicos: selección a priori

| objid | julian_date | ra | dec | redshift | u_filter | g_filter | r_filter | i_filter | z_filter | x_filter | camcol | field | plate | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6739133 | 53230 | 300.841762 | 76.5112824 | -0.00021952 | 19.19619 | 17.83329 | 17.52225 | 17.40237 | 17.35182 | 40131269 | 1 | 106 | 1660 | STAR |
| 4696263 | 53230 | 300.730508 | 76.5517315 | -8.36E-06 | 21.65541 | 19.13715 | 17.92577 | 17.44741 | 17.15818 | 98026415 | 1 | 106 | 1660 | STAR |
| 1905890 | 53230 | 300.871382 | 76.5305704 | 9.59E-05 | 20.70867 | 19.20954 | 18.55966 | 18.24395 | 18.10117 | 20605683 | 1 | 106 | 1660 | STAR |
| 2621494 | 53240 | 300.317409 | 76.374746 | -0.00024726 | 22.88806 | 21.209 | 19.9056 | 19.33555 | 19.08966 | | 1 | 106 | 1661 | STAR |
| 3421446 | 53230 | 301.252332 | 76.3195196 | -0.0001315 | 17.82932 | 16.11081 | 15.39808 | 15.13612 | 15.00507 | | 1 | 107 | 1660 | STAR |
| 5165444 | 53230 | 301.458518 | 76.4267657 | -0.00028757 | 14.23474 | 14.45194 | 14.75188 | 14.99473 | 15.24296 | | 1 | 107 | 1660 | STAR |
| 8145190 | 53230 | 300.848872 | 76.3204622 | -0.0002666 | 21.69935 | 19.60817 | 18.18657 | 17.04788 | 16.44566 | | 1 | 107 | 1660 | STAR |
| 2732193 | 53240 | 301.319371 | 76.3542733 | -0.00028874 | 25.33987 | 22.16937 | 20.70235 | 20.12803 | 19.72791 | | 1 | 107 | 1661 | STAR |
| 3868589 | 53240 | 301.003481 | 76.4351085 | -0.00064419 | 21.77835 | 20.39711 | 19.72937 | 19.43841 | 19.2693 | | 1 | 107 | 1661 | STAR |
| 6845117 | 53230 | 302.163901 | 76.388992 | -0.0002625 | 20.26735 | 18.5646 | 17.85914 | 17.58408 | 17.43464 | | 1 | 108 | 1660 | STAR |
| 8465026 | 53230 | 302.174176 | 76.1933405 | -7.03E-05 | 21.54818 | 19.57721 | 18.43685 | 17.93544 | 17.62113 | | 1 | 109 | 1660 | STAR |
| 2009504 | 53240 | 302.531797 | 76.3260686 | -2.79E-05 | 23.76005 | 21.57068 | 20.1001 | 19.45527 | 19.15964 | | 1 | 109 | 1661 | STAR |
| 5734622 | 53230 | 302.53014 | 76.1376457 | -4.62E-05 | 21.58243 | 19.6114 | 18.55185 | 18.13491 | 17.91136 | | 1 | 110 | 1660 | STAR |
| 8124788 | 53230 | 302.920973 | 76.3215062 | 2.68E-05 | 20.61019 | 18.13416 | 16.82791 | 16.22867 | 15.85798 | | 1 | 110 | 1660 | STAR |
| 3893789 | 53240 | 303.625982 | 76.1874215 | -0.00020144 | 22.87485 | 21.75856 | 20.42418 | 19.75471 | 19.3883 | | 1 | 111 | 1661 | STAR |
| 8931301 | 53230 | 303.76272 | 76.2041805 | -5.20E-05 | 17.27701 | 15.49217 | 14.6868 | 14.34704 | 14.15624 | 92214002 | 1 | 111 | 1660 | STAR |
| 4973464 | 53230 | 303.073864 | 76.1124891 | -0.00015582 | 21.94913 | 19.14192 | 17.67034 | 17.05692 | 16.69741 | | 1 | 111 | 1660 | STAR |
| 7920184 | 53230 | 303.286843 | 76.0656598 | 6.70E-06 | 22.04301 | 19.78859 | 18.52149 | 18.03293 | 17.72046 | | 1 | 111 | 1660 | STAR |
| 2164794 | 53240 | 303.506589 | 76.077405 | -0.00023373 | 22.4063 | 20.3364 | 19.52508 | 19.19709 | 19.02766 | | 1 | 111 | 1661 | STAR |
| 5487131 | 53230 | 303.217422 | 76.1972257 | -9.25E-05 | 21.93067 | 20.09178 | 18.90102 | 18.33528 | 17.97431 | | 1 | 111 | 1660 | STAR |
| 5927640 | 53240 | 303.354846 | 76.0528235 | -0.00043435 | 24.16451 | 22.21421 | 20.74198 | 20.24756 | 19.84714 | | 1 | 111 | 1661 | STAR |
| 6297958 | 53230 | 303.852385 | 76.1738998 | -0.00023089 | 17.07861 | 15.71224 | 15.54605 | 15.49323 | 15.47547 | | 1 | 112 | 1660 | STAR |
| 1728668 | 53240 | 304.020676 | 76.1601276 | -0.00054259 | 21.39911 | 19.83503 | 19.43652 | 19.24213 | 19.18858 | | 1 | 112 | 1661 | STAR |
| 6343143 | 53230 | 303.983522 | 75.9810815 | -0.00014069 | 19.507 | 17.731 | 16.94764 | 16.65991 | 16.48766 | | 1 | 112 | 1660 | STAR |
| 2612032 | 53240 | 303.98435 | 76.0978291 | -0.00014655 | 22.84097 | 20.48384 | 19.03249 | 18.43891 | 18.0931 | | 1 | 112 | 1661 | STAR |
| 1405877 | 53240 | 303.941362 | 76.0801423 | -0.00029587 | 23.18969 | 20.9254 | 19.63644 | 19.03506 | 18.61127 | 20727395 | 1 | 112 | 1661 | STAR |
| 5771574 | 53240 | 304.152653 | 76.0944185 | -0.00019201 | 23.9366 | 21.53848 | 20.07633 | 19.51465 | 19.11936 | 14140728 | 1 | 112 | 1661 | STAR |
| 1410471 | 53230 | 304.531555 | 76.1320652 | -0.00022718 | 19.56016 | 18.1636 | 17.49738 | 17.2079 | 17.036 | 30894670 | 1 | 113 | 1660 | STAR |
| 3804784 | 53230 | 304.373259 | 75.9887762 | -0.00028019 | 21.98413 | 19.17101 | 17.96169 | 17.46724 | 17.22819 | | 1 | 113 | 1660 | STAR |
| 8551198 | 53230 | 304.602721 | 76.119609 | -0.00016839 | 21.64969 | 19.46481 | 18.30491 | 17.83545 | 17.54242 | | 1 | 113 | 1660 | STAR |
| 3498800 | 53240 | 304.458928 | 75.9339433 | -0.00031972 | 16.92797 | 15.63552 | 15.10081 | 14.89095 | 14.79415 | | 1 | 113 | 1661 | STAR |
| 9187177 | 53230 | 304.447666 | 76.0485287 | -1.22E-05 | 19.25177 | 16.60164 | 15.20536 | 14.54854 | 14.19345 | 92679336 | 1 | 113 | 1660 | STAR |
| 2518609 | 53230 | 304.535053 | 75.9621095 | -0.0003306 | 20.77139 | 19.33941 | 18.614 | 18.37783 | 18.19239 | 90707782 | 1 | 113 | 1660 | STAR |

Características

Clase

# Datos astronómicos: selección a priori

| objid | julian_date | ra | dec | redshift | u_filter | g_filter | r_filter | i_filter | z_filter | x_filter | camcol | field | plate | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 673913 | 53230 | 300.841762 | 76.5112824 | -0.00021952 | 19.19619 | 17.83329 | 17.52225 | 17.40237 | 17.35182 | 40131269 | 1 | 106 | 1660 | STAR |
| 469626 | 53230 | 300.730508 | 76.5517315 | -8.36E-06 | 21.65541 | 19.13715 | 17.92577 | 17.44741 | 17.15818 | 98026415 | 1 | 106 | 1660 | STAR |
| 190589 | 53230 | 300.871382 | 76.5305704 | 9.59E-05 | 20.70867 | 19.20954 | 18.55966 | 18.24395 | 18.10117 | 20605683 | 1 | 106 | 1660 | STAR |
| 262149 | 53240 | 300.317409 | 76.374746 | -0.00024726 | 22.88806 | 21.209 | 19.9056 | 19.33555 | 19.08966 | | 1 | 106 | 1661 | STAR |
| 342146 | 53230 | 301.252332 | 76.3195196 | -0.0001315 | 17.82932 | 16.11081 | 15.39808 | 15.13612 | 15.00507 | | 1 | 107 | 1660 | STAR |
| 516544 | 53230 | 301.458518 | 76.4267657 | -0.00028757 | 14.23474 | 14.45194 | 14.75188 | 14.99473 | 15.24296 | | 1 | 107 | 1660 | STAR |
| 814590 | 53230 | 300.848872 | 76.3204622 | -0.0002666 | 21.69935 | 19.60817 | 18.18657 | 17.04788 | 16.44566 | | 1 | 107 | 1660 | STAR |
| 273293 | 53240 | 301.319371 | 76.3542733 | -0.00028874 | 25.33987 | 22.16937 | 20.70235 | 20.12803 | 19.72791 | | 1 | 107 | 1661 | STAR |
| 386589 | 53240 | 301.003481 | 76.4351085 | -0.00064419 | 21.77835 | 20.39711 | 19.72937 | 19.43841 | 19.2693 | | 1 | 107 | 1661 | STAR |
| 584117 | 53230 | 302.163901 | 76.388992 | -0.0002625 | 20.26735 | 18.5646 | 17.85914 | 17.58408 | 17.43464 | | 1 | 108 | 1660 | STAR |
| 346026 | 53230 | 302.174176 | 76.1933405 | -7.03E-05 | 21.54818 | 19.57721 | 18.43685 | 17.93544 | 17.62113 | | 1 | 109 | 1660 | STAR |
| 109504 | 53240 | 302.531797 | 76.3260686 | -2.79E-05 | 23.76005 | 21.57068 | 20.1001 | 19.45527 | 19.15964 | | 1 | 109 | 1661 | STAR |
| 573622 | 53230 | 302.53014 | 76.1376457 | -4.62E-05 | 21.58243 | 19.6114 | 18.55185 | 18.13491 | 17.91136 | | 1 | 110 | 1660 | STAR |
| 814788 | 53230 | 302.920973 | 76.3215062 | 2.68E-05 | 20.61019 | 18.13416 | 16.82791 | 16.22867 | 15.85798 | | 1 | 110 | 1660 | STAR |
| 393789 | 53240 | 303.625982 | 76.1874215 | -0.00020144 | 22.87485 | 21.75856 | 20.42418 | 19.75471 | 19.3883 | | 1 | 111 | 1661 | STAR |
| 831301 | 53230 | 303.76272 | 76.2041805 | -5.20E-05 | 17.27701 | 15.49217 | 14.6868 | 14.34704 | 14.15624 | 92214002 | 1 | 111 | 1660 | STAR |
| 473464 | 53230 | 303.073864 | 76.1124891 | -0.00015582 | 21.94913 | 19.14192 | 17.67034 | 17.05692 | 16.69741 | | 1 | 111 | 1660 | STAR |
| 720184 | 53230 | 303.286843 | 76.0656598 | 6.70E-06 | 22.04301 | 19.78859 | 18.52149 | 18.03293 | 17.72046 | | 1 | 111 | 1660 | STAR |
| 264794 | 53240 | 303.506589 | 76.077405 | -0.00023373 | 22.4063 | 20.3364 | 19.52508 | 19.19709 | 19.02766 | | 1 | 111 | 1661 | STAR |
| 947131 | 53230 | 303.217422 | 76.1972257 | -9.25E-05 | 21.93067 | 20.09178 | 18.90102 | 18.33528 | 17.97431 | | 1 | 111 | 1660 | STAR |
| 927640 | 53240 | 303.354846 | 76.0528235 | -0.00043435 | 24.16451 | 22.21421 | 20.74198 | 20.24756 | 19.84714 | | 1 | 111 | 1661 | STAR |
| 529958 | 53230 | 303.852385 | 76.1738998 | -0.00023089 | 17.07861 | 15.71224 | 15.54605 | 15.49323 | 15.47547 | | 1 | 112 | 1660 | STAR |
| 172668 | 53240 | 304.020676 | 76.1601276 | -0.00054259 | 21.39911 | 19.83503 | 19.43652 | 19.24213 | 19.18858 | | 1 | 112 | 1661 | STAR |
| 634143 | 53230 | 303.983522 | 75.9810815 | -0.00014069 | 19.507 | 17.731 | 16.94764 | 16.65991 | 16.48766 | | 1 | 112 | 1660 | STAR |
| 261232 | 53240 | 303.98435 | 76.0978291 | -0.00014655 | 22.84097 | 20.48384 | 19.03249 | 18.43891 | 18.0931 | | 1 | 112 | 1661 | STAR |
| 1405877 | 53240 | 303.941362 | 76.0801423 | -0.00029587 | 23.18969 | 20.9254 | 19.63644 | 19.03506 | 18.61127 | 20727395 | 1 | 112 | 1661 | STAR |
| 577174 | 53240 | 304.152653 | 76.0944185 | -0.00019201 | 23.9366 | 21.53848 | 20.07633 | 19.51465 | 19.11936 | 14140728 | 1 | 112 | 1661 | STAR |
| 1410471 | 53230 | 304.531555 | 76.1320652 | -0.00022718 | 19.56016 | 18.1636 | 17.49738 | 17.2079 | 17.036 | 30894670 | 1 | 113 | 1660 | STAR |
| 3804784 | 53230 | 304.373259 | 75.9887762 | -0.00028019 | 21.98413 | 19.17101 | 17.96169 | 17.46724 | 17.22819 | | 1 | 113 | 1660 | STAR |
| 8551198 | 53230 | 304.602721 | 76.119609 | -0.00016839 | 21.64969 | 19.46481 | 18.30491 | 17.83545 | 17.54242 | | 1 | 113 | 1660 | STAR |
| 3498800 | 53240 | 304.458928 | 75.9339433 | -0.00031972 | 16.92797 | 15.63552 | 15.10081 | 14.89095 | 14.79415 | | 1 | 113 | 1661 | STAR |
| 9187177 | 53230 | 304.447666 | 76.0485287 | -1.22E-05 | 19.25177 | 16.60164 | 15.20536 | 14.54854 | 14.19345 | 92679336 | 1 | 113 | 1660 | STAR |
| 2518609 | 53230 | 304.535053 | 75.9621095 | -0.0003306 | 20.77139 | 19.33941 | 18.614 | 18.37783 | 18.19239 | 90707782 | 1 | 113 | 1660 | STAR |

Características

Clase

# DATOS

## ¿Qué hacemos con los datos faltantes (información incompleta)?

- Hay que ver dónde están los datos faltantes.
- Si una característica tiene la mayoría de los datos faltantes, generalmente lo mejor es eliminar esa característica.
- Si una característica tiene pocos datos faltantes, podemos:
  - Eliminar la fila donde falte esa característica.
  - Reemplazar el dato faltante por un promedio, un valor por defecto, o un aproximado según sea el caso.

# DATOS

## ¿Cómo le entregamos los datos a un modelo para entrenarlo?

- Hay que separar los datos en dos: una matriz de características y un vector de clases.
- La matriz de características, en inglés generalmente llamada *feature matrix*, *feature vector* o simplemente *features*, la denotamos con una "X".
- El vector de clases, en inglés generalmente llamado como *target vector* o simplemente *labels*, lo denotamos con una "y".
- En las librerías, los parámetros de las funciones generalmente usan estos nombres (X e y).

# Datos astronómicos: *features* y *labels*

| objid | julian_date | ra | dec | redshift | u_filter | g_filter | r_filter | i_filter | z_filter | x_filter | camcol | field | plate | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6739133 | 53230 | 300.841762 | 76.5112824 | -0.00021952 | 19.19619 | 17.83329 | 17.52225 | 17.40237 | 17.35182 | 40131269 | 1 | 106 | 1660 | STAR |
| 4696263 | 53230 | 300.730508 | 76.5517315 | -8.36E-06 | 21.65541 | 19.13715 | 17.92577 | 17.44741 | 17.15818 | 98026415 | 1 | 106 | 1660 | STAR |
| 1905890 | 53230 | 300.871382 | 76.5305704 | 9.59E-05 | 20.70867 | 19.20954 | 18.55966 | 18.24395 | 18.10117 | 20605683 | 1 | 106 | 1660 | STAR |
| 2621494 | 53240 | 300.317409 | 76.374746 | -0.00024726 | 22.88806 | 21.209 | 19.9056 | 19.33555 | 19.08966 |  | 1 | 106 | 1661 | STAR |
| 3421446 | 53230 | 301.252332 | 76.3195196 | -0.0001315 | 17.82932 | 16.11081 | 15.39808 | 15.13612 | 15.00507 |  | 1 | 107 | 1660 | STAR |
| 5165444 | 53230 | 301.458518 | 76.4267657 | -0.00028757 | 14.23474 | 14.45194 | 14.75188 | 14.99473 | 15.24296 |  | 1 | 107 | 1660 | STAR |
| 8145190 | 53230 | 300.848872 | 76.3204622 | -0.0002666 | 21.69935 | 19.60817 | 18.18657 | 17.04788 | 16.44566 |  | 1 | 107 | 1660 | STAR |
| 2732193 | 53240 | 301.319371 | 76.3542733 | -0.00028874 | 25.33987 | 22.16937 | 20.70235 | 20.12803 | 19.72791 |  | 1 | 107 | 1661 | STAR |
| 3868589 | 53240 | 301.003481 | 76.4351085 | -0.00064419 | 21.77835 | 20.39711 | 19.72937 | 19.43841 | 19.2693 |  | 1 | 107 | 1661 | STAR |
| 6845117 | 53230 | 302.163901 | 76.388992 | -0.0002625 | 20.26735 | 18.5646 | 17.85914 | 17.58408 | 17.43464 |  | 1 | 108 | 1660 | STAR |
| 8465026 | 53230 | 302.174176 | 76.1933405 | -7.03E-05 | 21.54818 | 19.57721 | 18.43685 | 17.93544 | 17.62113 |  | 1 | 109 | 1660 | STAR |
| 2009504 | 53240 | 302.531797 | 76.3260686 | -2.79E-05 | 23.76005 | 21.57068 | 20.1001 | 19.45527 | 19.15964 |  | 1 | 109 | 1661 | STAR |
| 5734622 | 53230 | 302.53014 | 76.1376457 | -4.62E-05 | 21.58243 | 19.6114 | 18.55185 | 18.13491 | 17.91136 |  | 1 | 110 | 1660 | STAR |
| 8124788 | 53230 | 302.920973 | 76.3215062 | 2.68E-05 | 20.61019 | 18.13416 | 16.82791 | 16.22867 | 15.85798 |  | 1 | 110 | 1660 | STAR |
| 3893789 | 53240 | 303.625982 | 76.1874215 | -0.00020144 | 22.87485 | 21.75856 | 20.42418 | 19.75471 | 19.3883 |  | 1 | 111 | 1661 | STAR |
| 8931301 | 53230 | 303.76272 | 76.2041805 | -5.20E-05 | 17.27701 | 15.49217 | 14.6868 | 14.34704 | 14.15624 | 92214002 | 1 | 111 | 1660 | STAR |
| 4973464 | 53230 | 303.073864 | 76.1124891 | -0.00015582 | 21.94913 | 19.14192 | 17.67034 | 17.05692 | 16.69741 |  | 1 | 111 | 1660 | STAR |
| 7920184 | 53230 | 303.286843 | 76.0656598 | 6.70E-06 | 22.04301 | 19.78859 | 18.52149 | 18.03293 | 17.72046 |  | 1 | 111 | 1660 | STAR |
| 2164794 | 53240 | 303.506589 | 76.077405 | -0.00023373 | 22.4063 | 20.3364 | 19.52508 | 19.19709 | 19.02766 |  | 1 | 111 | 1661 | STAR |
| 5487131 | 53230 | 303.217422 | 76.1972257 | -9.25E-05 | 21.93067 | 20.09178 | 18.90102 | 18.33528 | 17.97431 |  | 1 | 111 | 1660 | STAR |
| 5927640 | 53240 | 303.354846 | 76.0528235 | -0.00043435 | 24.16451 | 22.21421 | 20.74198 | 20.24756 | 19.84714 |  | 1 | 111 | 1661 | STAR |
| 6297958 | 53230 | 303.852385 | 76.1738998 | -0.00023089 | 17.07861 | 15.71224 | 15.54605 | 15.49323 | 15.47547 |  | 1 | 112 | 1660 | STAR |
| 1728668 | 53240 | 304.020676 | 76.1601276 | -0.00054259 | 21.39911 | 19.83503 | 19.43652 | 19.24213 | 19.18858 |  | 1 | 112 | 1661 | STAR |
| 6343143 | 53230 | 303.983522 | 75.9810815 | -0.00014069 | 19.507 | 17.731 | 16.94764 | 16.65991 | 16.48766 |  | 1 | 112 | 1660 | STAR |
| 2612032 | 53240 | 303.98435 | 76.0978291 | -0.00014655 | 22.84097 | 20.48384 | 19.03249 | 18.43891 | 18.0931 |  | 1 | 112 | 1661 | STAR |
| 1405877 | 53240 | 303.941362 | 76.0801423 | -0.00029587 | 23.18969 | 20.9254 | 19.63644 | 19.03506 | 18.61127 | 20727395 | 1 | 112 | 1661 | STAR |
| 5771574 | 53240 | 304.152653 | 76.0944185 | -0.00019201 | 23.9366 | 21.53848 | 20.07633 | 19.51465 | 19.11936 | 14140728 | 1 | 112 | 1661 | STAR |
| 1410471 | 53230 | 304.531555 | 76.1320652 | -0.00022718 | 19.56016 | 18.1636 | 17.49738 | 17.2079 | 17.036 | 30894670 | 1 | 113 | 1660 | STAR |
| 3804784 | 53230 | 304.373259 | 75.9887762 | -0.00028019 | 21.98413 | 19.17101 | 17.96169 | 17.46724 | 17.22819 |  | 1 | 113 | 1660 | STAR |
| 8551198 | 53230 | 304.602721 | 76.119609 | -0.00016839 | 21.64969 | 19.46481 | 18.30491 | 17.83545 | 17.54242 |  | 1 | 113 | 1660 | STAR |
| 3498800 | 53240 | 304.458928 | 75.9339433 | -0.00031972 | 16.92797 | 15.63552 | 15.10081 | 14.89095 | 14.79415 |  | 1 | 113 | 1661 | STAR |
| 9187177 | 53230 | 304.447666 | 76.0485287 | -1.22E-05 | 19.25177 | 16.60164 | 15.20536 | 14.54854 | 14.19345 | 92679336 | 1 | 113 | 1660 | STAR |
| 2518609 | 53230 | 304.535053 | 75.9621095 | -0.0003306 | 20.77139 | 19.33941 | 18.614 | 18.37783 | 18.19239 | 90707782 | 1 | 113 | 1660 | STAR |

Matriz de características

Vector de clases

# Datos astronómicos: *features* y *labels*

| objid | julian_date | ra | dec | redshift | u_filter | g_filter | r_filter | i_filter | z_filter | x_filter | camcol | field | plate | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6739133 | 53230 | 300.841762 | 76.5112824 | -0.00021952 | 19.19619 | 17.83329 | 17.52225 | 17.40237 | 17.35182 | 40131269 | 1 | 106 | 1660 | STAR |
| 4696263 | 53230 | 300.730508 | 76.5517315 | -8.36E-06 | 21.65541 | 19.13715 | 17.92577 | 17.44741 | 17.15818 | 98026415 | 1 | 106 | 1660 | STAR |
| 1905890 | 53230 | 300.871382 | 76.5305704 | 9.59E-05 | 20.70867 | 19.20954 | 18.55966 | 18.24395 | 18.10117 | 20605683 | 1 | 106 | 1660 | STAR |
| 2621494 | 53240 | 300.317409 | 76.374746 | -0.00024726 | 22.88806 | 21.209 | 19.9056 | 19.33555 | 19.08966 |  | 1 | 106 | 1661 | STAR |
| 3421446 | 53230 | 301.252332 | 76.3195196 | -0.0001315 | 17.82932 | 16.11081 | 15.39808 | 15.13612 | 15.00507 |  | 1 | 107 | 1660 | STAR |
| 5165444 | 53230 | 301.458518 | 76.4267657 | -0.00028757 | 14.23474 | 14.45194 | 14.75188 | 14.99473 | 15.24296 |  | 1 | 107 | 1660 | STAR |
| 8145190 | 53230 | 300.848872 | 76.3204622 | -0.0002666 | 21.69935 | 19.60817 | 18.18657 | 17.04788 | 16.44566 |  | 1 | 107 | 1660 | STAR |
| 2732193 | 53240 | 301.319371 | 76.3542733 | -0.00028874 | 25.33987 | 22.16937 | 20.70235 | 20.12803 | 19.72791 |  | 1 | 107 | 1661 | STAR |
| 3868589 | 53240 | 301.003481 | 76.4351085 | -0.00064419 | 21.77835 | 20.39711 | 19.72937 | 19.43841 | 19.2693 |  | 1 | 107 | 1661 | STAR |
| 6845117 | 53230 | 302.163901 | 76.388992 | -0.0002625 | 20.26735 | 18.5646 | 17.85914 | 17.58408 | 17.43464 |  | 1 | 108 | 1660 | STAR |
| 8465026 | 53230 | 302.174176 | 76.1933405 | -7.03E-05 | 21.54818 | 19.57721 | 18.43685 | 17.93544 | 17.62113 |  | 1 | 109 | 1660 | STAR |
| 2009504 | 53240 | 302.531797 | 76.3260686 | -2.79E-05 | 23.76005 | 21.57068 | 20.1001 | 19.45527 | 19.15964 |  | 1 | 109 | 1661 | STAR |
| 5734622 | 53230 | 302.53014 | 76.1376457 | -4.62E-05 | 21.58243 | 19.6114 | 18.55185 | 18.13491 | 17.91136 |  | 1 | 110 | 1660 | STAR |
| 8124788 | 53230 | 302.920973 | 76.3215062 | 2.68E-05 | 20.61019 | 18.13416 | 16.82791 | 16.22867 | 15.85798 |  | 1 | 110 | 1660 | STAR |
| 3893789 | 53240 | 303.625982 | 76.1874215 | -0.00020144 | 22.87485 | 21.75856 | 20.42418 | 19.75471 | 19.3883 |  | 1 | 111 | 1661 | STAR |
| 8931301 | 53230 | 303.76272 | 76.2041805 | -5.20E-05 | 17.27701 | 15.49217 | 14.6868 | 14.34704 | 14.15624 | 92214002 | 1 | 111 | 1660 | STAR |
| 4973464 | 53230 | 303.073864 | 76.1124891 | -0.00015582 | 21.94913 | 19.14192 | 17.67034 | 17.05692 | 16.69741 |  | 1 | 111 | 1660 | STAR |
| 7920184 | 53230 | 303.286843 | 76.0656598 | 6.70E-06 | 22.04301 | 19.78859 | 18.52149 | 18.03293 | 17.72046 |  | 1 | 111 | 1660 | STAR |
| 2164794 | 53240 | 303.506589 | 76.077405 | -0.00023373 | 22.4063 | 20.3364 | 19.52508 | 19.19709 | 19.02766 |  | 1 | 111 | 1661 | STAR |
| 5487131 | 53230 | 303.217422 | 76.1972257 | -9.25E-05 | 21.93067 | 20.09178 | 18.90102 | 18.33528 | 17.97431 |  | 1 | 111 | 1660 | STAR |
| 5927640 | 53240 | 303.354846 | 76.0528235 | -0.00043435 | 24.16451 | 22.21421 | 20.74198 | 20.24756 | 19.84714 |  | 1 | 111 | 1661 | STAR |
| 6297958 | 53230 | 303.852385 | 76.1738998 | -0.00023089 | 17.07861 | 15.71224 | 15.54605 | 15.49323 | 15.47547 |  | 1 | 112 | 1660 | STAR |
| 1728668 | 53240 | 304.020676 | 76.1601276 | -0.00054259 | 21.39911 | 19.83503 | 19.43652 | 19.24213 | 19.18858 |  | 1 | 112 | 1661 | STAR |
| 6343143 | 53230 | 303.983522 | 75.9810815 | -0.00014069 | 19.507 | 17.731 | 16.94764 | 16.65991 | 16.48766 |  | 1 | 112 | 1660 | STAR |
| 2612032 | 53240 | 303.98435 | 76.0978291 | -0.00014655 | 22.84097 | 20.48384 | 19.03249 | 18.43891 | 18.0931 |  | 1 | 112 | 1661 | STAR |
| 1405877 | 53240 | 303.941362 | 76.0801423 | -0.00029587 | 23.18969 | 20.9254 | 19.63644 | 19.03506 | 18.61127 | 20727395 | 1 | 112 | 1661 | STAR |
| 5771574 | 53240 | 304.152653 | 76.0944185 | -0.00019201 | 23.9366 | 21.53848 | 20.07633 | 19.51465 | 19.11936 | 14140728 | 1 | 112 | 1661 | STAR |
| 1410471 | 53230 | 304.531555 | 76.1320652 | -0.00022718 | 19.56016 | 18.1636 | 17.49738 | 17.2079 | 17.036 | 30894670 | 1 | 113 | 1660 | STAR |
| 3804784 | 53230 | 304.373259 | 75.9887762 | -0.00028019 | 21.98413 | 19.17101 | 17.96169 | 17.46724 | 17.22819 |  | 1 | 113 | 1660 | STAR |
| 8551198 | 53230 | 304.602721 | 76.119609 | -0.00016839 | 21.64969 | 19.46481 | 18.30491 | 17.83545 | 17.54242 |  | 1 | 113 | 1660 | STAR |
| 3498800 | 53240 | 304.458928 | 75.9339433 | -0.00031972 | 16.92797 | 15.63552 | 15.10081 | 14.89095 | 14.79415 |  | 1 | 113 | 1661 | STAR |
| 9187177 | 53230 | 304.447666 | 76.0485287 | -1.22E-05 | 19.25177 | 16.60164 | 15.20536 | 14.54854 | 14.19345 | 92679336 | 1 | 113 | 1660 | STAR |
| 2518609 | 53230 | 304.535053 | 75.9621095 | -0.0003306 | 20.77139 | 19.33941 | 18.614 | 18.37783 | 18.19239 | 90707782 | 1 | 113 | 1660 | STAR |

X

y

# DATOS

## ¿Usamos todos los datos para entrenar al modelo?

- Generalmente no. Si usamos todos los datos después no nos quedan datos para probarlo cómo funciona el modelo.
- Por esta razón dividimos el dataset en dos: set de entrenamiento y set de pruebas.
- Mientras más datos usemos para entrenar el modelo mejor, por lo tanto el set de entrenamiento debe ser lo más grande posible.
- Por otra parte queremos dejar una cantidad representativa de datos para probar.
- Dependiendo del volumen de datos, una recomendación es 70% - 30%

# Datos astronómicos: entrenamiento y pruebas

X_train

| objid | julian_date | ra | dec | redshift | u_filter | g_filter | r_filter | i_filter | z_filter | x_filter | camcol | field | plate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6739133 | 53230 | 300.841762 | 76.5112824 | -0.00021952 | 19.19619 | 17.83329 | 17.52225 | 17.40237 | 17.35182 | 40131269 | 1 | 106 | 1660 |
| 4696263 | 53230 | 300.730508 | 76.5517315 | -8.36E-06 | 21.65541 | 19.13715 | 17.92577 | 17.44741 | 17.15818 | 98026415 | 1 | 106 | 1660 |
| 1905890 | 53230 | 300.871382 | 76.5305704 | 9.59E-05 | 20.70867 | 19.20954 | 18.55966 | 18.24395 | 18.10117 | 20605683 | 1 | 106 | 1660 |
| 2621494 | 53240 | 300.317409 | 76.374746 | -0.00024726 | 22.88806 | 21.209 | 19.9056 | 19.33555 | 19.08966 | | 1 | 106 | 1661 |
| 3421446 | 53230 | 301.252332 | 76.3195196 | -0.0001315 | 17.82932 | 16.11081 | 15.39808 | 15.13612 | 15.00507 | | 1 | 107 | 1660 |
| 5165444 | 53230 | 301.458518 | 76.4267657 | -0.00028757 | 14.23474 | 14.45194 | 14.75188 | 14.99473 | 15.24296 | | 1 | 107 | 1660 |
| 8145190 | 53230 | 300.848872 | 76.3204622 | -0.0002666 | 21.69935 | 19.60817 | 18.18657 | 17.04788 | 16.44566 | | 1 | 107 | 1660 |
| 2732193 | 53240 | 301.319371 | 76.3542733 | -0.00028874 | 25.33987 | 22.16937 | 20.70235 | 20.12803 | 19.72791 | | 1 | 107 | 1661 |
| 3868589 | 53240 | 301.003481 | 76.4351085 | -0.00064419 | 21.77835 | 20.39711 | 19.72937 | 19.43841 | 19.2693 | | 1 | 107 | 1661 |
| 6845117 | 53230 | 302.163901 | 76.388992 | -0.0002625 | 20.26735 | 18.5646 | 17.85914 | 17.58408 | 17.43464 | | 1 | 108 | 1660 |
| 8465026 | 53230 | 302.174176 | 76.1933405 | -7.03E-05 | 21.54818 | 19.57721 | 18.43685 | 17.93544 | 17.62113 | | 1 | 109 | 1660 |
| 2009504 | 53240 | 302.531797 | 76.3260686 | -2.79E-05 | 23.76005 | 21.57068 | 20.1001 | 19.45527 | 19.15964 | | 1 | 109 | 1661 |
| 5734622 | 53230 | 302.53014 | 76.1376457 | -4.62E-05 | 21.58243 | 19.6114 | 18.55185 | 18.13491 | 17.91136 | | 1 | 110 | 1660 |
| 8124788 | 53230 | 302.920973 | 76.3215062 | 2.68E-05 | 20.61019 | 18.13416 | 16.82791 | 16.22867 | 15.85798 | | 1 | 110 | 1660 |
| 3893789 | 53240 | 303.625982 | 76.1874215 | -0.00020144 | 22.87485 | 21.75856 | 20.42418 | 19.75471 | 19.3883 | | 1 | 111 | 1661 |
| 8931301 | 53230 | 303.76272 | 76.2041805 | -5.20E-05 | 17.27701 | 15.49217 | 14.6868 | 14.34704 | 14.15624 | 92214002 | 1 | 111 | 1660 |
| 4973464 | 53230 | 303.073864 | 76.1124891 | -0.00015582 | 21.94913 | 19.14192 | 17.67034 | 17.05692 | 16.69741 | | 1 | 111 | 1660 |
| 7920184 | 53230 | 303.286843 | 76.0656598 | 6.70E-06 | 22.04301 | 19.78859 | 18.52149 | 18.03293 | 17.72046 | | 1 | 111 | 1660 |
| 2164794 | 53240 | 303.506589 | 76.077405 | -0.00023373 | 22.4063 | 20.3364 | 19.52508 | 19.19709 | 19.02766 | | 1 | 111 | 1661 |
| 5487131 | 53230 | 303.217422 | 76.1972257 | -9.25E-05 | 21.93067 | 20.09178 | 18.90102 | 18.33528 | 17.97431 | | 1 | 111 | 1660 |
| 5927640 | 53240 | 303.354846 | 76.0528235 | -0.00043435 | 24.16451 | 22.21421 | 20.74198 | 20.24756 | 19.84714 | | 1 | 111 | 1661 |
| 6297958 | 53230 | 303.852385 | 76.1738998 | -0.00023089 | 17.07861 | 15.71224 | 15.54605 | 15.49323 | 15.47547 | | 1 | 112 | 1660 |
| 1728668 | 53240 | 304.020676 | 76.1601276 | -0.00054259 | 21.39911 | 19.83503 | 19.43652 | 19.24213 | 19.18858 | | 1 | 112 | 1661 |
| 6343143 | 53230 | 303.983522 | 75.9810815 | -0.00014069 | 19.507 | 17.731 | 16.94764 | 16.65991 | 16.48766 | | 1 | 112 | 1660 |
| 2612032 | 53240 | 303.98435 | 76.0978291 | -0.00014655 | 22.84097 | 20.48384 | 19.03249 | 18.43891 | 18.0931 | | 1 | 112 | 1661 |
| 1405877 | 53240 | 303.941362 | 76.0801423 | -0.00029587 | 23.18969 | 20.9254 | 19.63644 | 19.03506 | 18.61127 | 20727395 | 1 | 112 | 1661 |
| 5771574 | 53240 | 304.152653 | 76.0944185 | -0.00019201 | 23.9366 | 21.53848 | 20.07633 | 19.51465 | 19.11936 | 14140728 | 1 | 112 | 1661 |

y_train

| class |
|---|
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |

X_test

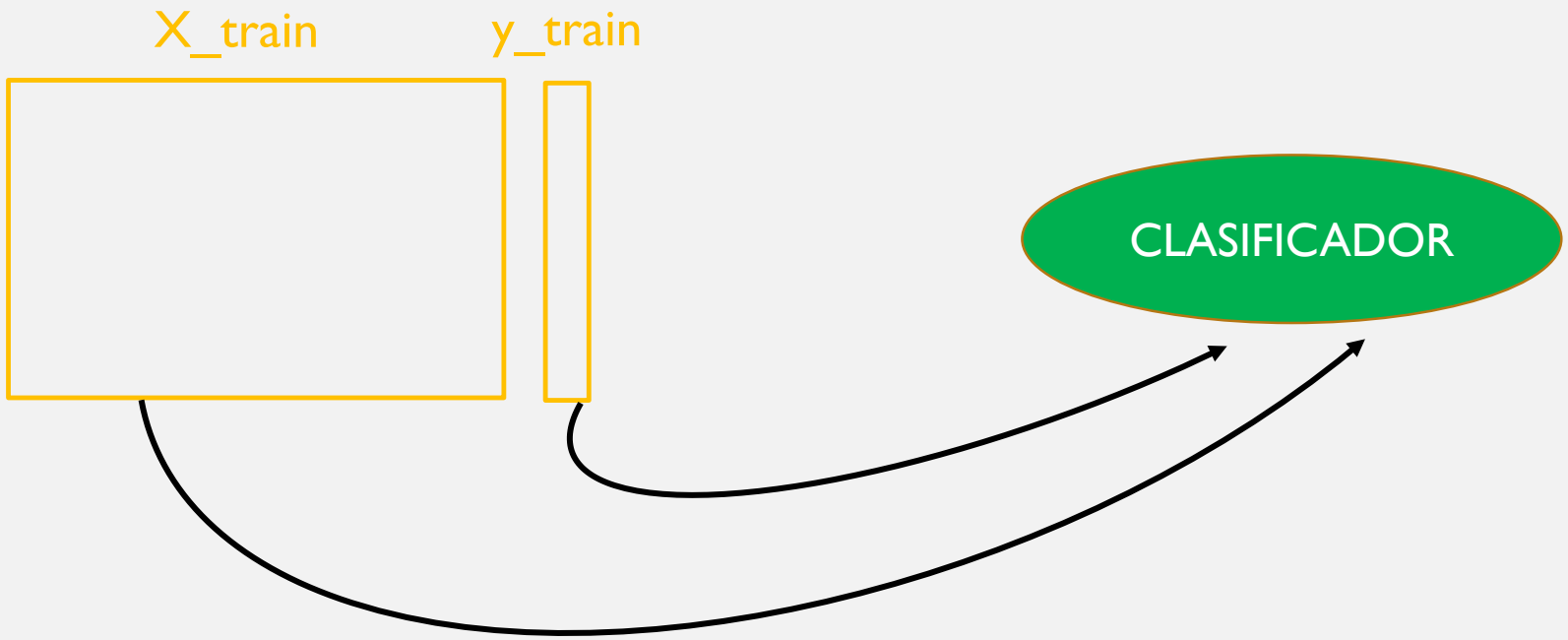
| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1410471 | 53230 | 304.531555 | 76.1320652 | -0.00022718 | 19.56016 | 18.1636 | 17.49738 | 17.2079 | 17.036 | 30894670 | 1 | 113 | 1660 |
| 3804784 | 53230 | 304.373259 | 75.9887762 | -0.00028019 | 21.98413 | 19.17101 | 17.96169 | 17.46724 | 17.22819 | | 1 | 113 | 1660 |
| 8551198 | 53230 | 304.602721 | 76.119609 | -0.00016839 | 21.64969 | 19.46481 | 18.30491 | 17.83545 | 17.54242 | | 1 | 113 | 1660 |
| 3498800 | 53240 | 304.458928 | 75.9339433 | -0.00031972 | 16.92797 | 15.63552 | 15.10081 | 14.89095 | 14.79415 | | 1 | 113 | 1661 |
| 9187177 | 53230 | 304.447666 | 76.0485287 | -1.22E-05 | 19.25177 | 16.60164 | 15.20536 | 14.54854 | 14.19345 | 92679336 | 1 | 113 | 1660 |
| 2518609 | 53230 | 304.535053 | 75.9621095 | -0.0003306 | 20.77139 | 19.33941 | 18.614 | 18.37783 | 18.19239 | 90707782 | 1 | 113 | 1660 |

y_test

| class |
|---|
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |
| STAR |

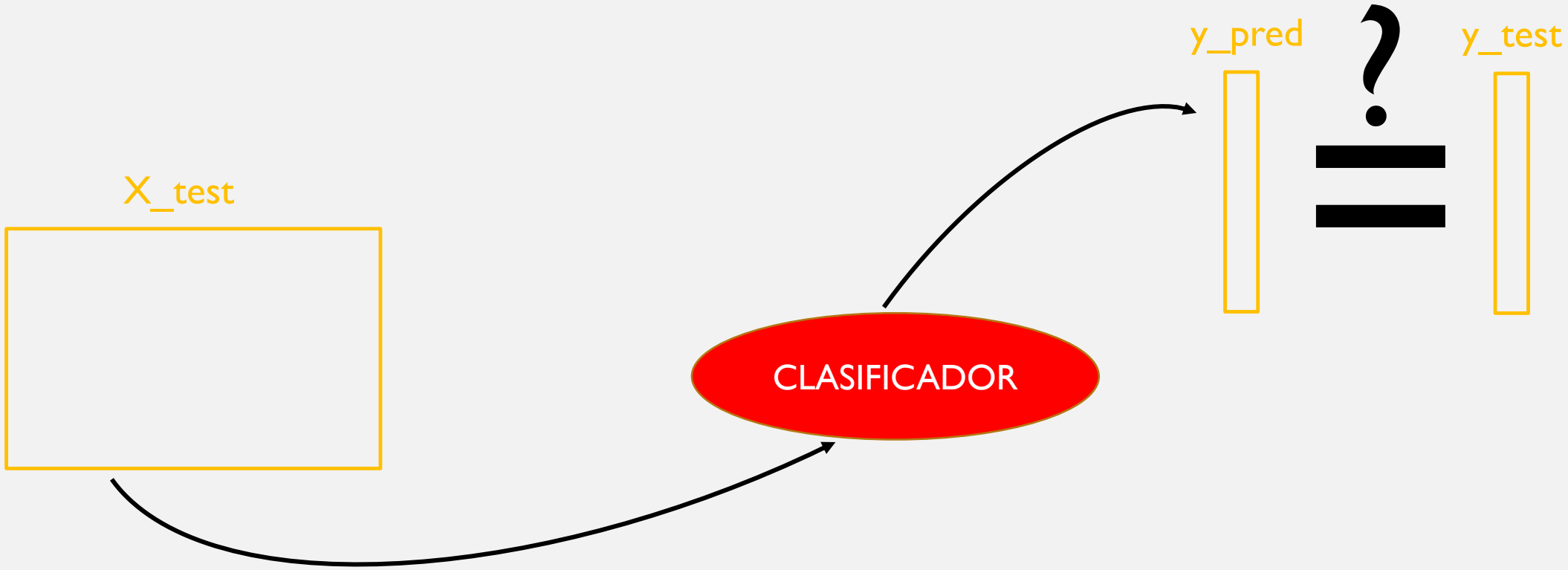
# Etapa de aprendizaje en un clasificador

Fase de aprendizaje



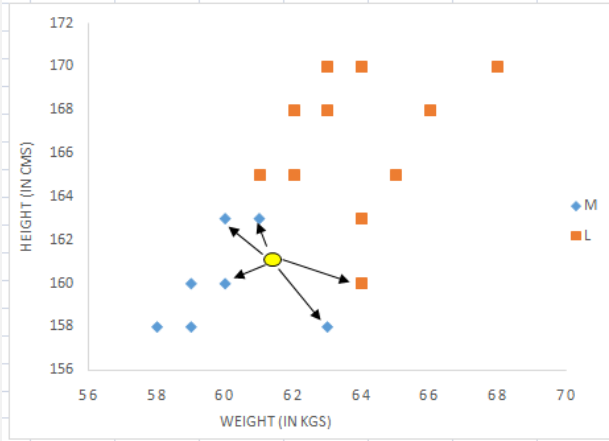
# Etapa de pruebas en un clasificador

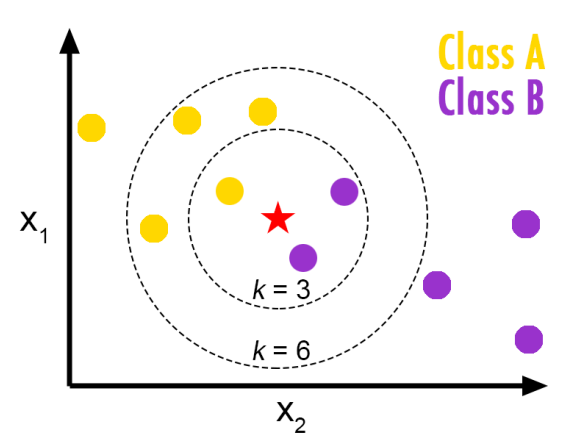
Fase de pruebas



# K-NEAREST NEIGHBOURS

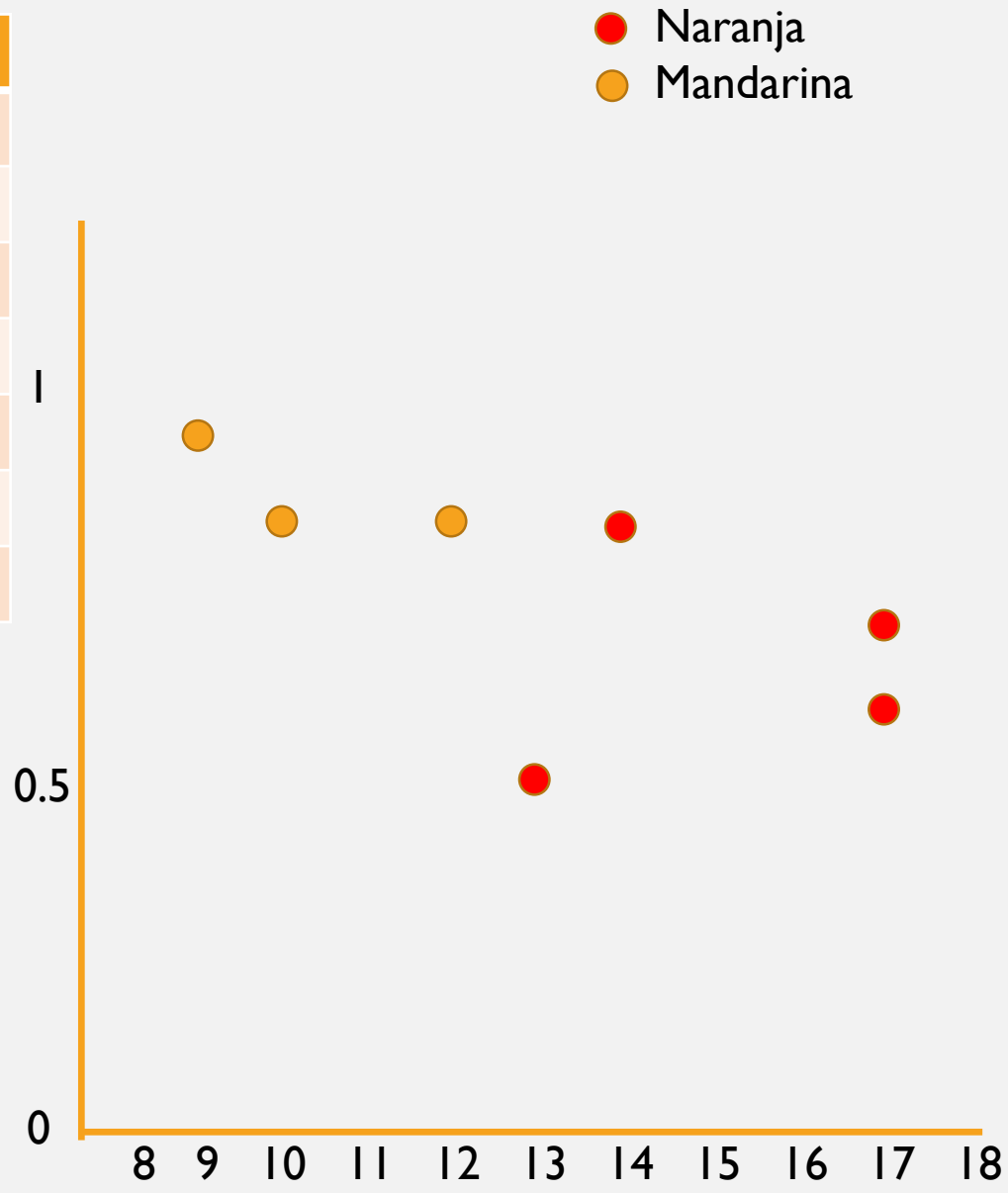
- Es una de las técnicas más sencillas de clasificación.
- Se basa en la "cercanía" que tienen las muestras.



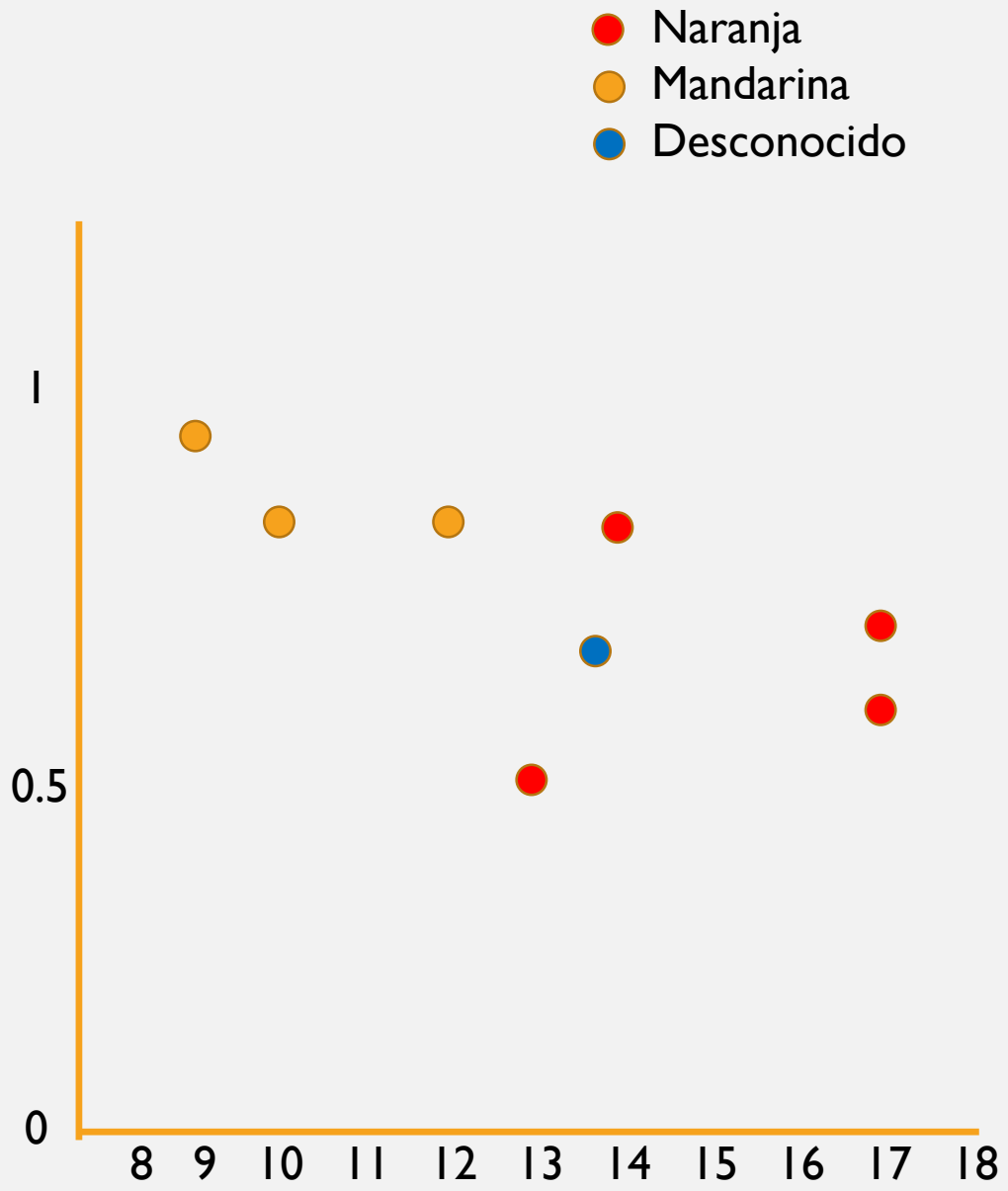



# Ejemplo KNN

| Diámetro (cm) | Intensidad de color | Clase |
|---|---|---|
| 10 | 0.8 | Mandarina |
| 12 | 0.8 | Mandarina |
| 9 | 0.9 | Mandarina |
| 17 | 0.6 | Naranja |
| 17 | 0.7 | Naranja |
| 14 | 0.8 | Naranja |
| 13 | 0.5 | Naranja |



# Ejemplo KNN


Naranja
Mandarina
Desconocido
1
0.5
0
8
9
10
11
12
13
14
15
16
17
18

# Ejemplo KNN

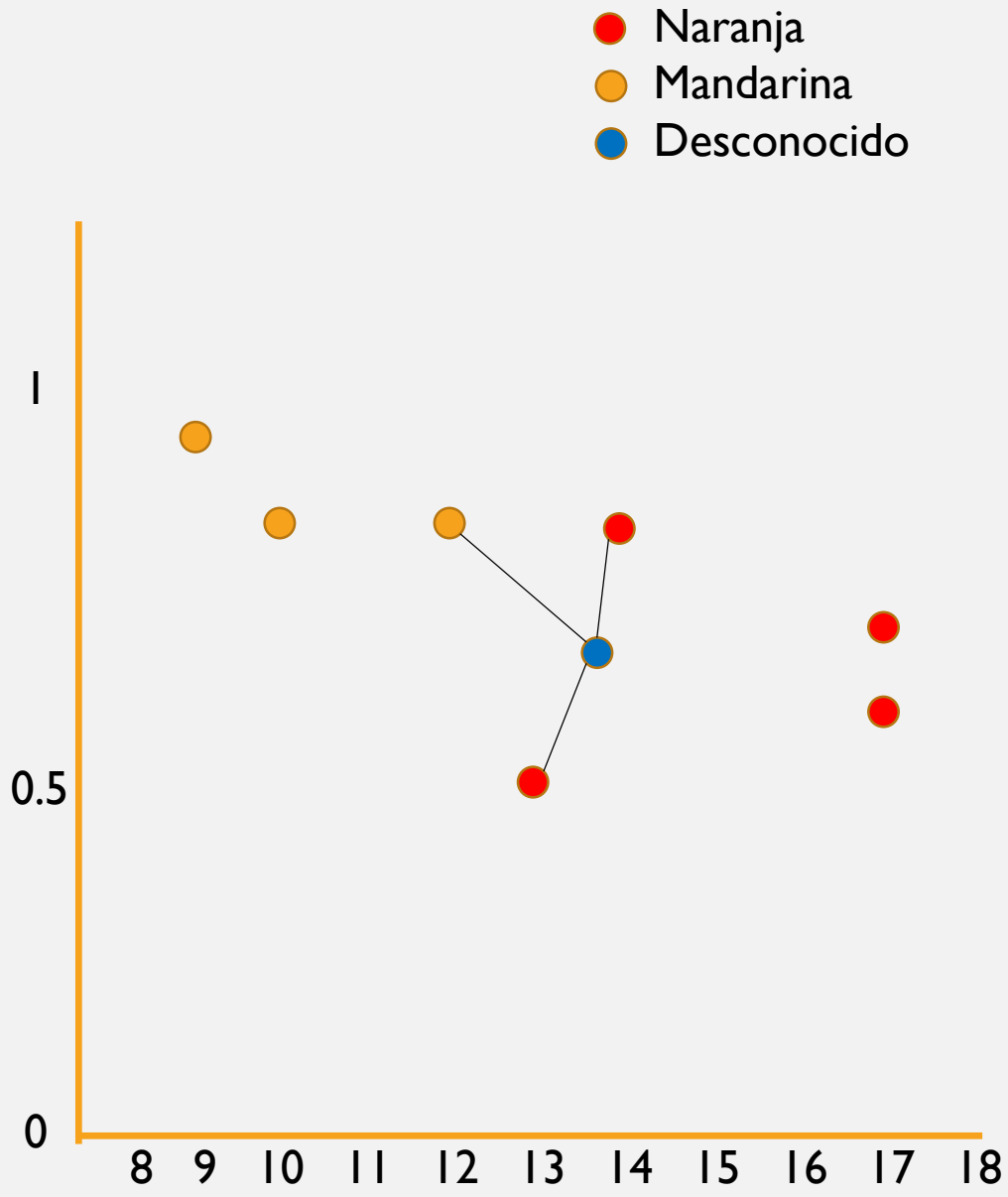
## Parámetro **K** del modelo

## Ejemplo: K = 3

Los vecinos son:
- 2 naranjas
- 1 mandarina

Por lo tanto el valor desconocido es una naranja

¿Y si hay empate? ¿Qué valores son buenos para K?
¿Que pasaría si K = 7?



# Ejemplo KNN

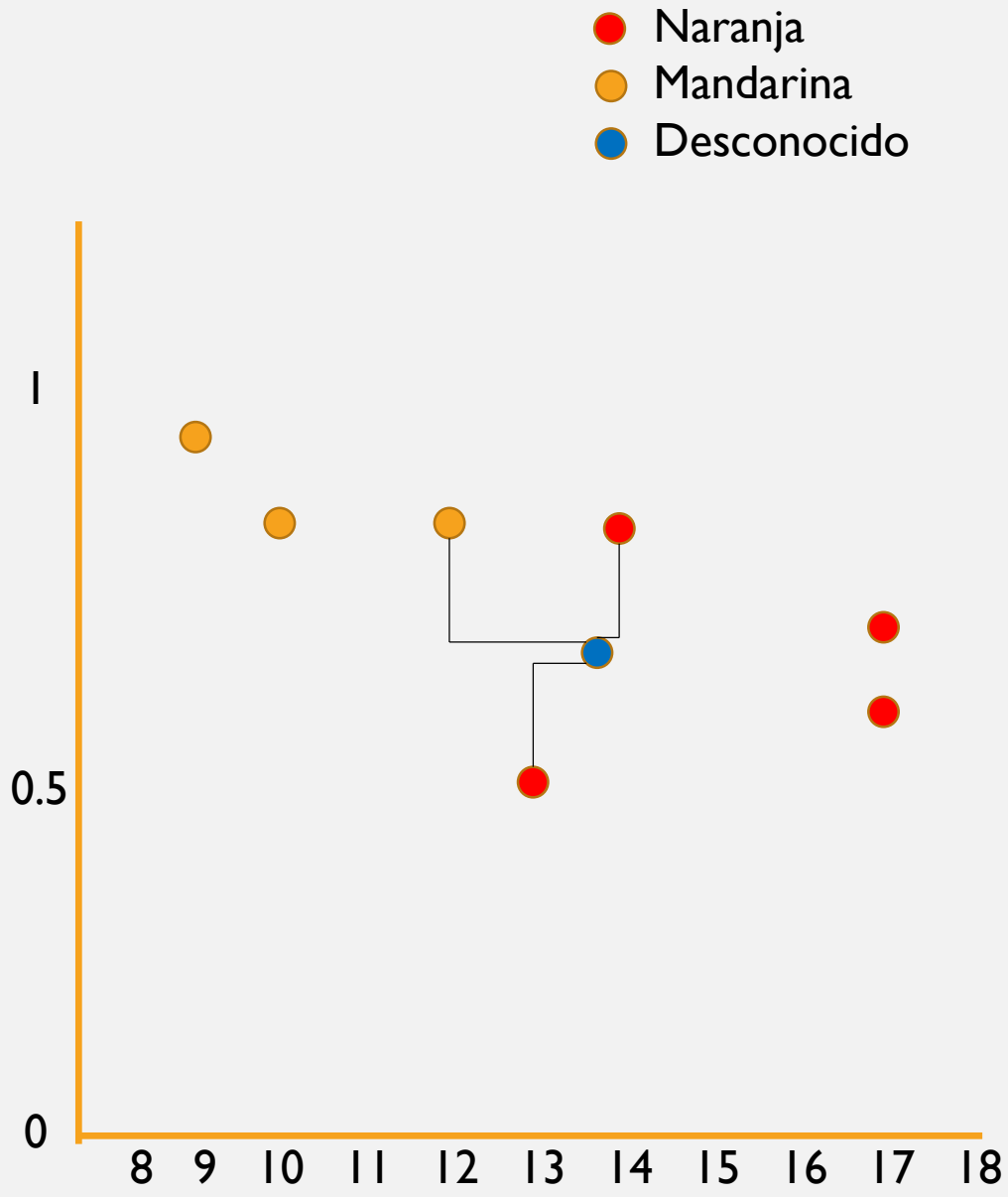
## Parámetro **Distancia** del modelo

## Ejemplo: manhattan

Según el tipo de distancia podría cambiar el resultado de la predicción.

La más usada: euclidiana. Se pueden inventar distintas fórmulas para las distancias.

Si uno de mis atributos es el color de pelo, ¿cómo represento eso en mis datos? ¿cómo mido la distancia entre dos valores?



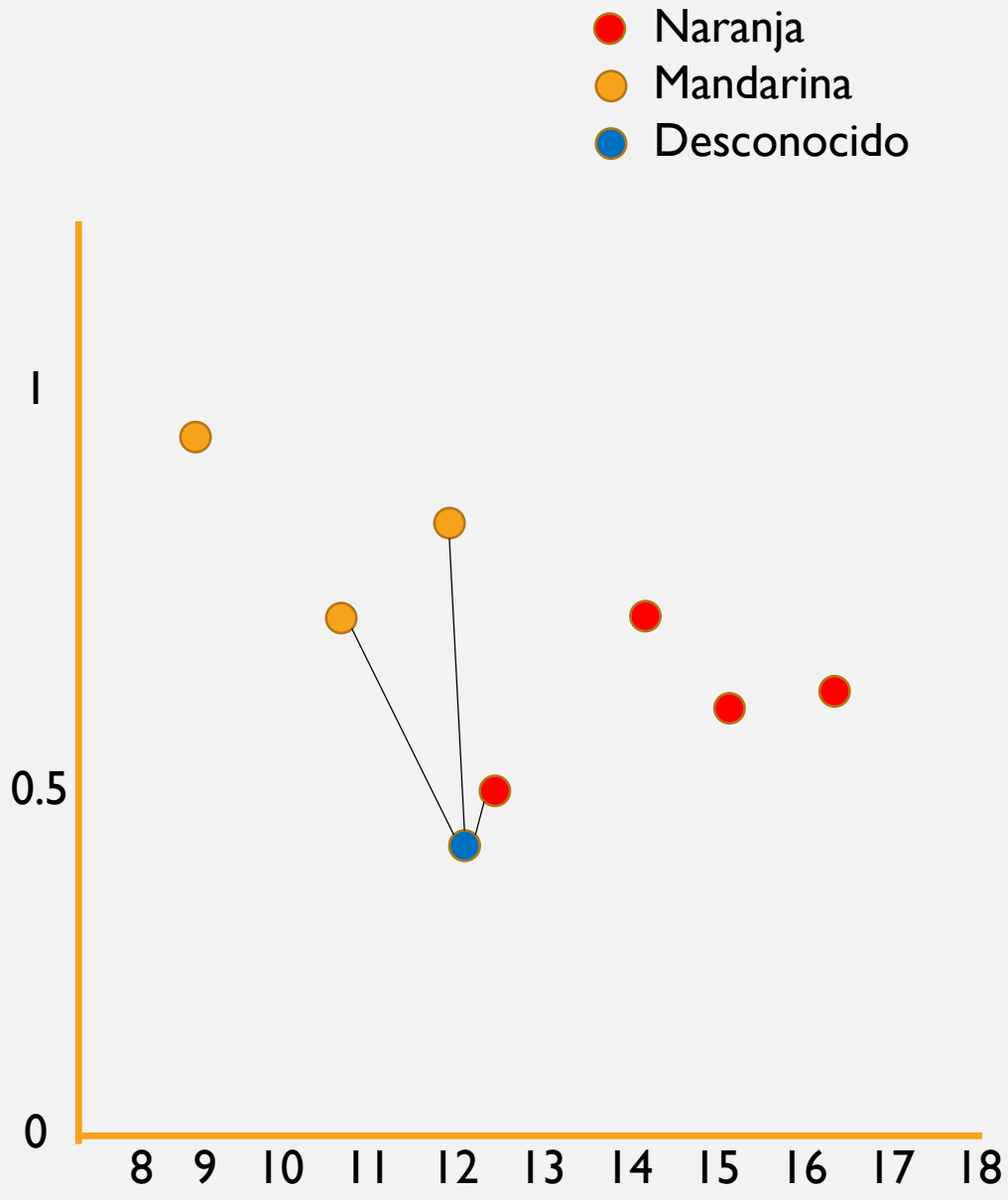
# Ejemplo KNN

## Parámetro **Peso** del modelo

## Ejemplo: inverso a distancia

Pondera cada vecino según su distancia (mientras más cerca, mayor importancia).

El método más común es con peso uniforme (todos los vecinos tienen igual importancia). El problema de empate que se presenta generalmente con peso uniforme, no suele ocurrir aquí.



# DATOS

## ¿Cómo influye la magnitud de los números en KNN?

- Dos características podrían moverse en rangos de números muy distintos.
  - La característica A podría tener valores entre 0 y 1
  - La característica B podría tener valores entre 10000 y 100000
- Esto afecta el algoritmo a la hora de medir las distancias.
  - Una diferencia de 0.5 en la característica A es bastante, pero en la característica B esa diferencia es marginal.
- Las características con valores más grandes van a tener mayor importancia a la hora de calcular la distancia entre dos puntos.

# DATOS

## Normalización

- Para solucionar este problema podemos normalizar los datos.
- Consiste en llevar las características a una misma escala.
- Generalmente se normaliza cada característica por separado.
  - ¿Por qué no tiene mucho sentido normalizar las filas?
  - ¿Que producen los *outliers* en una característica a la hora de normalizar?
- Existen distintas formas de normalizar: max, min-max, L1, L2

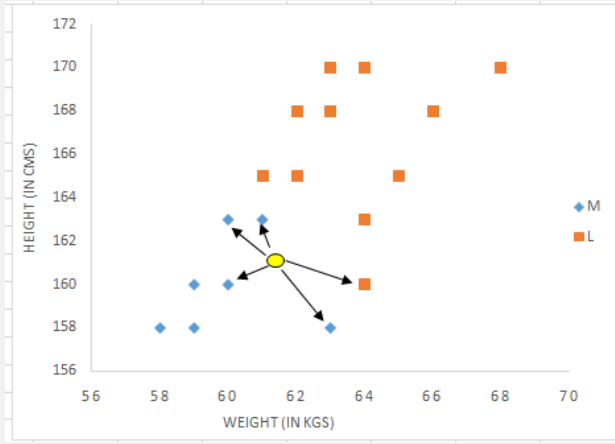
# Ejemplo Normalización: max

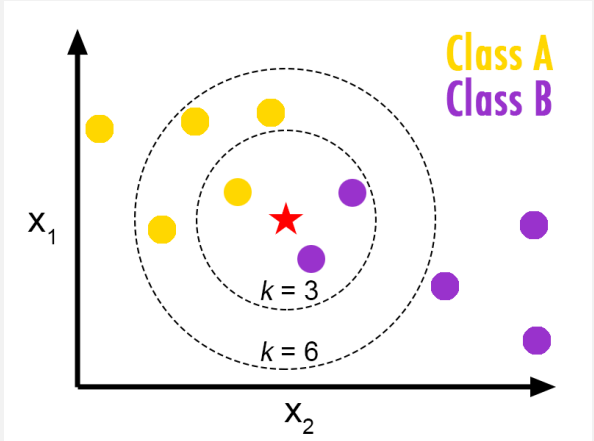
| Diámetro (cm) | Intensidad de color | Clase |
|---|---|---|
| 10 | 0.8 | Mandarina |
| 12 | 0.8 | Mandarina |
| 9 | 0.9 | Mandarina |
| 17 | 0.6 | Naranja |
| 17 | 0.7 | Naranja |
| 14 | 0.8 | Naranja |
| 13 | 0.5 | Naranja |

| Diámetro (cm) | Intensidad de color | Clase |
|---|---|---|
| 0.58 | 0.88 | Mandarina |
| 0.7 | 0.88 | Mandarina |
| 0.53 | 1 | Mandarina |
| 1 | 0.66 | Naranja |
| 1 | 0.77 | Naranja |
| 0.82 | 0.88 | Naranja |
| 0.76 | 0.55 | Naranja |

# K-NEAREST NEIGHBOURS

- Se almacenan los datos etiquetados en un espacio de N dimensiones.
- Para clasificar una instancia nueva, vemos a qué clase corresponden los K vecinos más cercanos.
- Generalmente se usa distancia euclidiana.
- El peso asignado a cada vecino puede ser uniforme o puede estar ponderado por su distancia.
- Es importante normalizar los datos.



# DECISION TREES

- A partir de los datos etiquetados se construye un árbol.
- Para clasificar una instancia nueva vamos avanzando por el árbol (según los atributos de esa instancia) hasta llegar a un nodo hoja que nos dice su clase.
- Se puede construir el árbol bajo distintas reglas y criterios.
- Los atributos con valores numéricos continuos deben tratarse por intervalos.
- Es importante entrenar con datos variados.
- Parámetros importantes: profundidad máxima del árbol y muestras mínimas para ramificar.

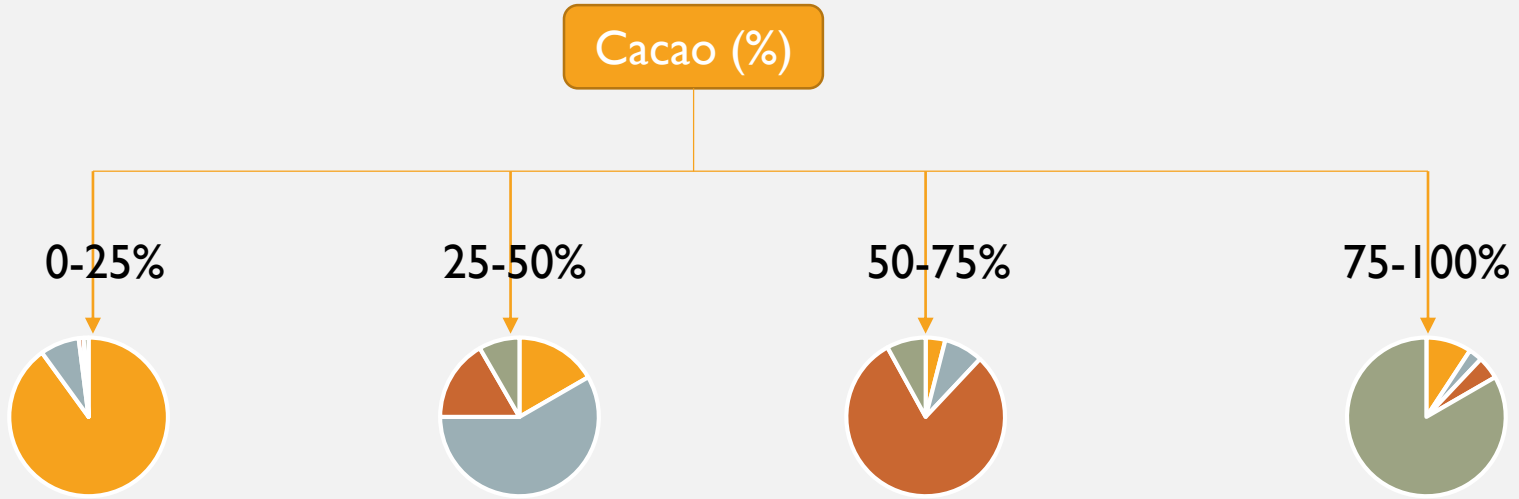
| Cacao (%) | Leche (%) | Color | Precio (por 100gr) | Calidad |
|---|---|---|---|---|
| 30 | 50 | Café neutro | 4312 | Muy bueno |
| 50 | 20 | Café oscuro | 4602 | Regular |
| 80 | 5 | Café oscuro | 8160 | Muy bueno |
| 20 | 40 | Café neutro | 2569 | Malo |
| 10 | 60 | Café claro | 1420 | Malo |
| 10 | 70 | Café claro | 1032 | Bueno |
| 30 | 20 | Café neutro | 4926 | Bueno |
| 60 | 10 | Café oscuro | 8741 | Regular |
| 55 | 5 | Café oscuro | 8423 | Muy bueno |
| 62 | 5 | Café oscuro | 9851 | Muy bueno |
| 20 | 40 | Café neutro | 4563 | Malo |
| 5 | 75 | Café claro | 5102 | Regular |
| 20 | 20 | Café neutro | 2036 | Malo |
| 15 | 30 | Café claro | 2471 | Malo |
| 20 | 30 | Café neutro | 3625 | Regular |
| 10 | 60 | Café claro | 1359 | Malo |
| 90 | 2 | Café oscuro | 10465 | Muy bueno |
| 30 | 20 | Café neutro | 2512 | Regular |

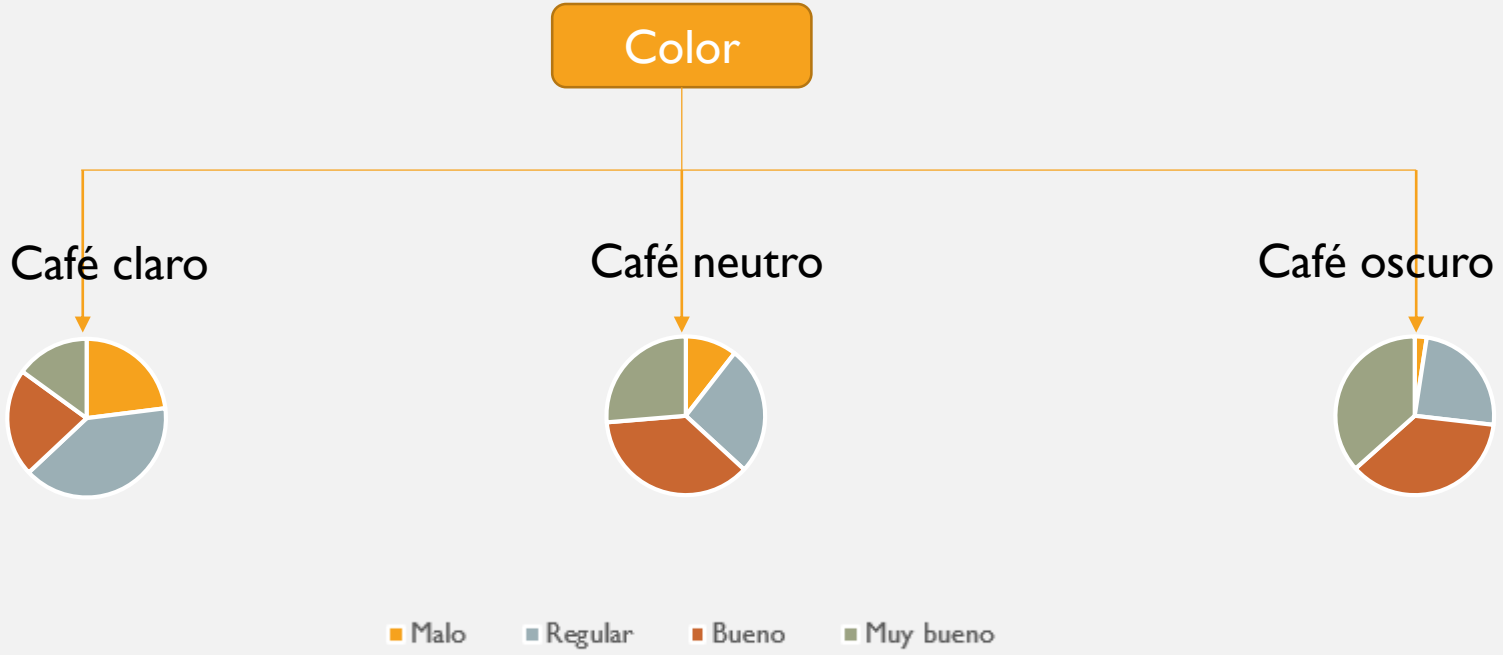
¿Qué atributo tiene mayor relación con la calidad?

¿Hay forma de medir/cuantificar esta relación?

¿En qué parte del árbol situamos a los que mejor separan las clases?

¿Hasta que punto seguimos ramificando el árbol?



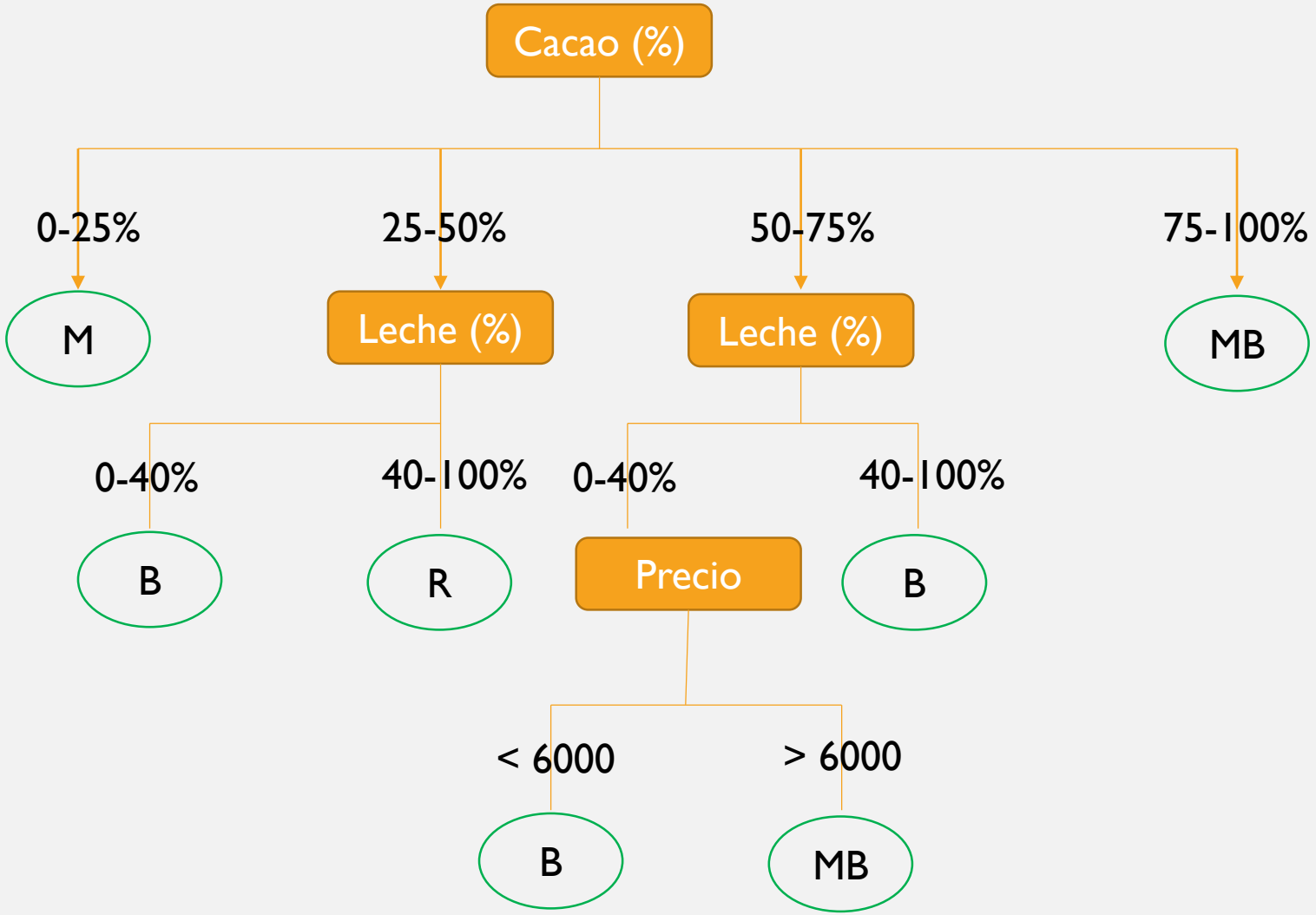



| Cacao (%) | Leche (%) | Color | Precio (por 100gr) | Calidad |
|---|---|---|---|---|
| 30 | 50 | Café neutro | 4312 | Muy bueno |
| 50 | 20 | Café oscuro | 4602 | Regular |
| 80 | 5 | Café oscuro | 8160 | Muy bueno |
| 20 | 40 | Café neutro | 2569 | Malo |
| 10 | 60 | Café claro | 1420 | Malo |
| 10 | 70 | Café claro | 1032 | Bueno |
| 30 | 20 | Café neutro | 4926 | Bueno |
| 60 | 10 | Café oscuro | 8741 | Regular |
| 55 | 5 | Café oscuro | 8423 | Muy bueno |
| 62 | 5 | Café oscuro | 9851 | Muy bueno |
| 20 | 40 | Café neutro | 4563 | Malo |
| 5 | 75 | Café claro | 5102 | Regular |
| 20 | 20 | Café neutro | 2036 | Malo |
| 15 | 30 | Café claro | 2471 | Malo |
| 20 | 30 | Café neutro | 3625 | Regular |
| 10 | 60 | Café claro | 1359 | Malo |
| 90 | 2 | Café oscuro | 10465 | Muy bueno |
| 30 | 20 | Café neutro | 2512 | Regular |

Para cuantificar qué atributos separan mejor las clases podemos usar la entropía o la impureza de Gini

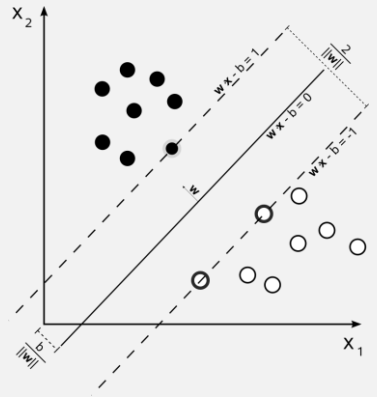
Situamos más cerca del nodo raíz a aquellos atributos que mejor separan las clases.

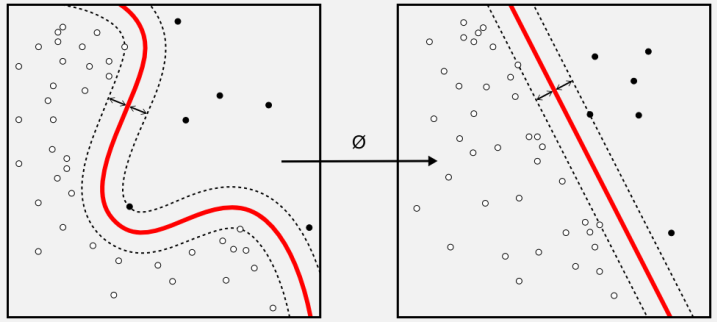
Mientras más profundo estamos en el árbol, menos ejemplos estaremos considerando. No es buena idea tomar la decisión cuando tenemos pocos ejemplos.



# SUPPORT VECTOR MACHINES

- Se encuentra un hiperplano que separa las distintas clases lo mejor posible.
- Para clasificar una instancia nueva, vemos en qué lado del hiperplano se encuentra.
- Se usan funciones de kernel para aumentar la dimensionalidad y así lograr una separación lineal.



# CLASIFICADORES

## ¿Cómo medimos su rendimiento?

- Vimos que existen distintos clasificadores, y que para entrenarlos debemos entregarles datos etiquetados.
- Al finalizar el entrenamiento nos gustaría darles un dato nuevo (que no tiene etiqueta) y ver si son capaces de clasificarlo correctamente.
- Por esta razón el conjunto de datos etiquetados los dividimos en dos: el set de entrenamiento y el set de prueba.

## SCORE O ACCURACY

Nos dice qué porcentaje de los datos que probamos fueron clasificados de manera correcta.

## MATRIZ DE CONFUSIÓN

Versión detallada del score, por clase.

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | **Iris-setosa** | **Iris-versicolor** | **Iris-virginica** | **Σ** |
| Actual | **Iris-setosa** | 100.0 % | 0.0 % | 0.0 % | **50** |
| | **Iris-versicolor** | 0.0 % | 88.7 % | 6.4 % | **50** |
| | **Iris-virginica** | 0.0 % | 11.3 % | 93.6 % | **50** |
| | **Σ** | **50** | **53** | **47** | **150** |

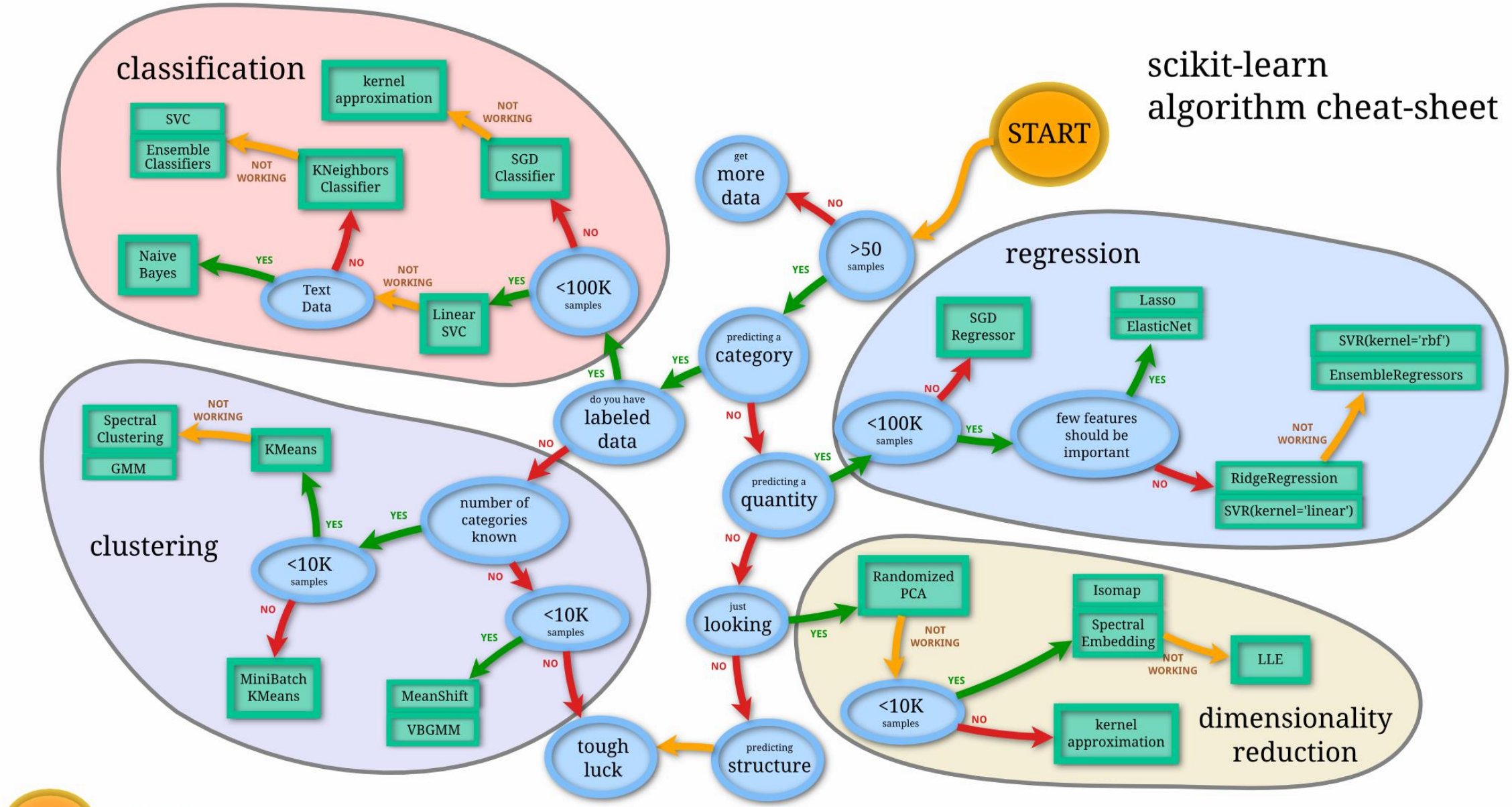
- Es importante que el set de pruebas tenga varios ejemplos de cada clase, y a su vez, la cantidad de ejemplos sea homogénea en todas las clases.
  - Se puede representar como valor numérico o como porcentaje.
- La MC permite ver qué clases son identificadas mejor y peor.
- La MC permite ver qué clases suelen confundirse más.

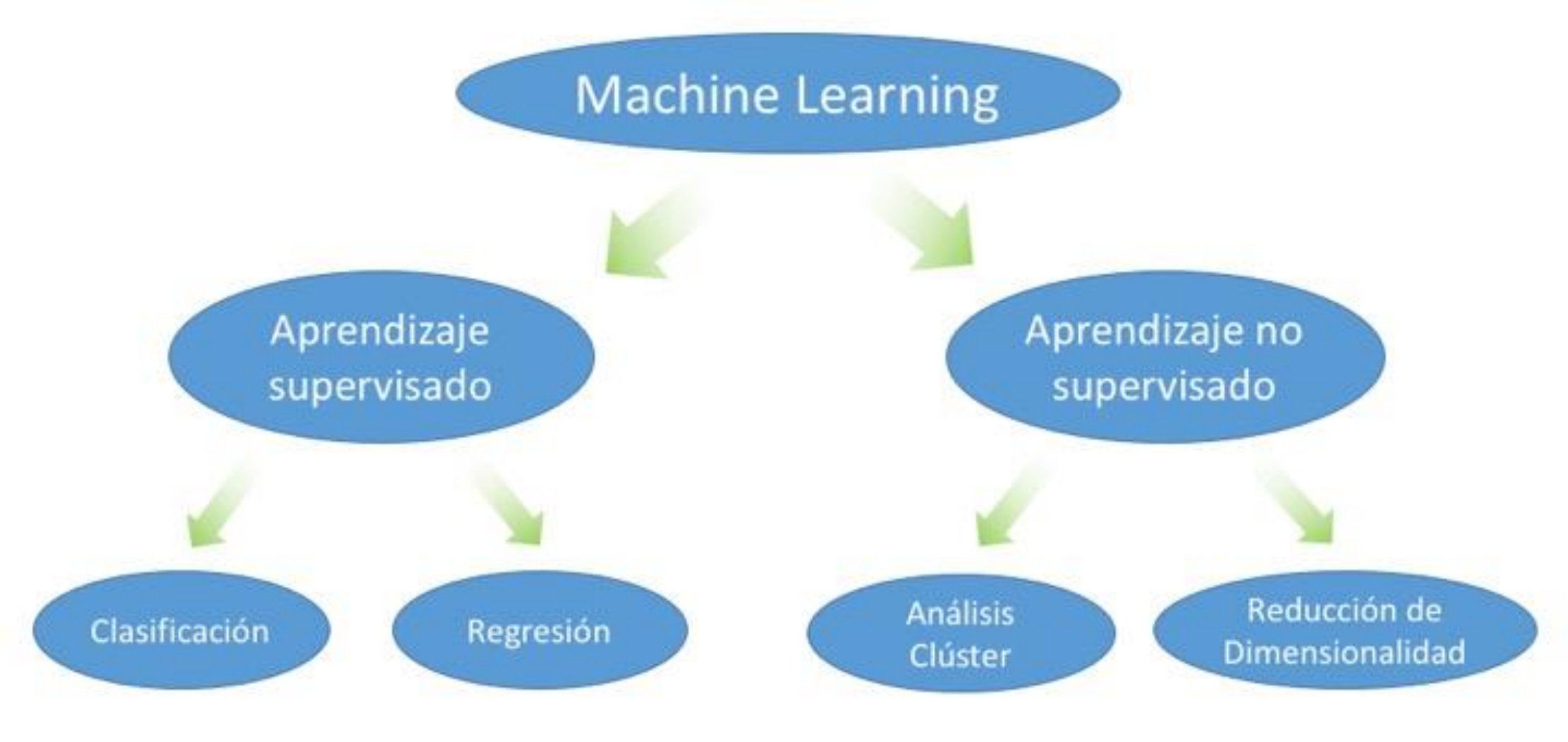
# CLASIFICADORES

## ¿Qué otras métricas existen?

- Funciones de pérdida (cuánto varía el valor predicho del valor real):
  - Logistic loss o crossentropy loss
- Métricas específicas para clasificación binaria (dos clases).
  - Precision
  - Recall (Sensitivity)
  - F1 Score
  - F-Beta Score
  - AUC-ROC

# OTRAS ÁREAS DE MACHINE LEARNING


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K samples
NO
get more data
NO
>50 samples
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
NO
RidgeRegression
SVR(kernel='linear')
predicting a category
YES
NO
do you have labeled data
YES
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
NO
<10K samples
NO
MiniBatch KMeans
<10K samples
YES
NO
MeanShift
VBGMM
just looking
YES
NO
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
tough luck
predicting structure
Back
scikit learn




Conoce las clases

No conoce las clases

# REDUCCIÓN DE DIMENSIONALIDAD

# REDUCCIÓN DE DIMENSIONALIDAD

## ¿Para qué sirve?

- Usamos técnicas de reducción de dimensionalidad cuando queremos visualizar los datos.
- Cuando tenemos datos en muchas dimensiones no es posible ver su distribución en el espacio, por lo tanto hacemos una reducción de dimensiones.
- Generalmente hacemos la reducción a 2 o 3 dimensiones.
- Al hacer la reducción hay una pérdida de información asociada.
- Otra utilidad es trabajar la misma información con menos datos.

# REDUCCIÓN DE DIMENSIONALIDAD

## ¿Qué técnicas existen?

- PCA
  - Hace una transformación lineal de la matriz de características.
  - Ordena las columnas de mayor a menor varianza.
  - Se seleccionan las primeras columnas.

- t-SNE

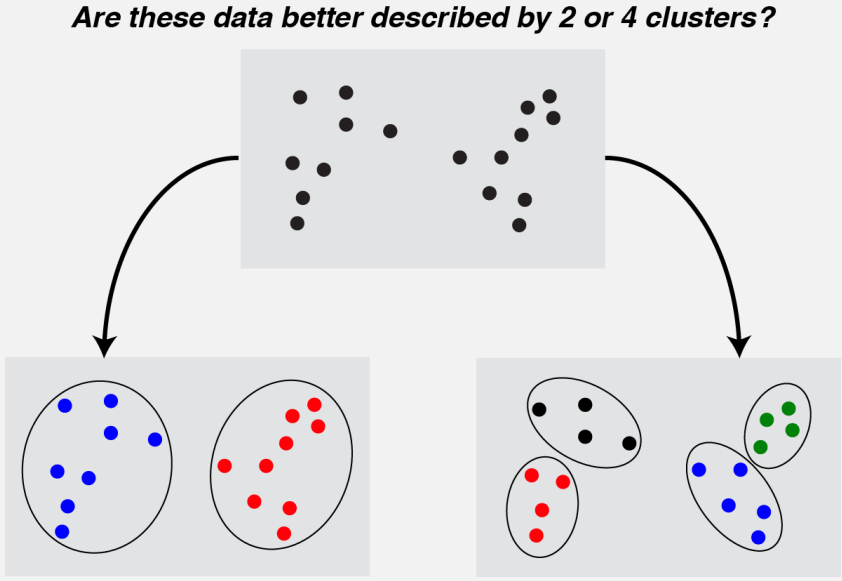
# CLUSTERING

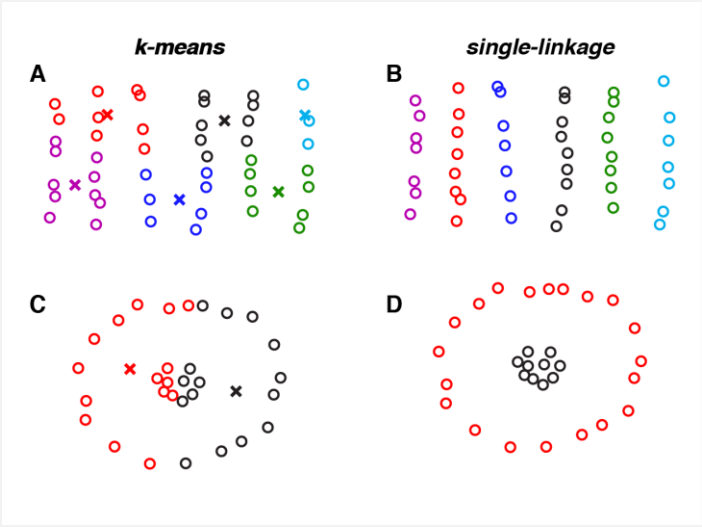
# CLUSTERING

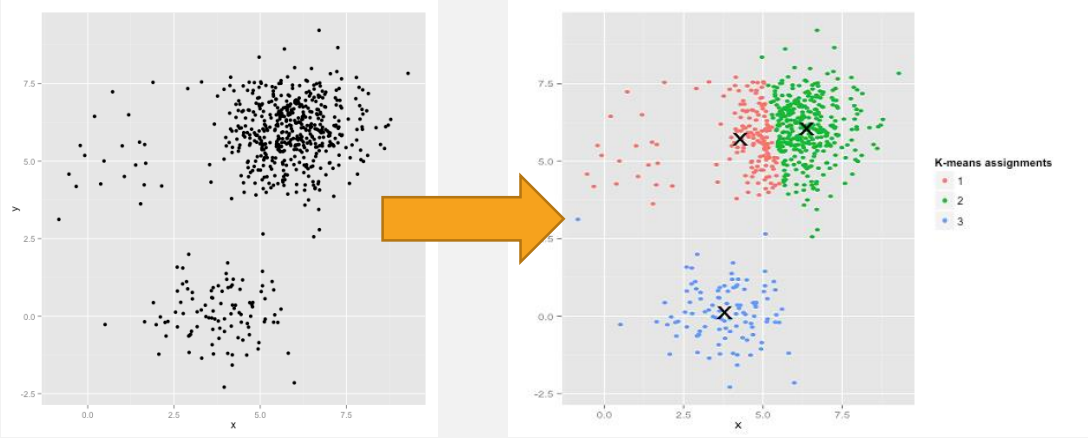
## ¿Para qué sirve?

- Usamos técnicas de clustering cuando tenemos datos no etiquetados y queremos encontrar las clases o categorías existentes.
- Si lo vemos desde una perspectiva visual puede ser un problema fácil de resolver para un humano, pero suele ser un problema difícil.
- Dependiendo del caso particular pueden haber varias soluciones posibles, es un tema de perspectiva.
- Aunque generalmente visualizamos las técnicas de clustering en 2D, estos pueden ser aplicados en una cantidad arbitraria de dimensiones.

- Los computadores no “ven” como los humanos, solo tienen una “lista de puntos”.
- Una solución podría encontrar más clases que otras. En general en estos problemas no se sabe el número de clases por lo tanto ambas opciones podrían ser correctas.
- Hay algoritmos que funcionan bien en algunos casos y otros que funcionan mejor en otros. Algo que puede parecer obvio para los humanos puede no serlo para un algoritmo.



# CLUSTERING

## ¿Qué algoritmos existen?

- Basados en conectividad
  - Single-linkage, complete-linkage, UPGMA
- Basados en centroide
  - K-Means y variaciones
- Basados en distribución
  - Gaussian mixture model
- Basados en densidad
  - Meanshift, DBSCAN y variaciones

# CLUSTERING

## ¿Cómo medimos su rendimiento?

- El desempeño de un algoritmo de clustering puede ser bueno o malo dependiendo de lo que se busca, por lo tanto es algo bastante subjetivo.
- Existen métricas que intentan medir con un puntaje el rendimiento del algoritmo. Se conocen como métricas de evaluación interna:
  - Índice de Davies-Bouldin, Índice de Dunn, Coeficiente de Silhouette
- Generalmente, estas métricas buscan ven qué tanto se parecen los puntos de un mismo cluster y qué tanto difieren puntos de clusters distintos. Esto hace que evalúen mejor cierto tipo de algoritmos.

# CLUSTERING

## ¿Cómo medimos su rendimiento?

- También existen métricas de evaluación externa donde una se entrega una solución al problema de clustering y luego se compara qué tan parecida es la respuesta entregada por el algoritmo.

# CONCEPTOS BÁSICOS DE MACHINE LEARNING

Martín De la Fuente