CONCEPTOS BÁSICOS DE MACHINE LEARNING

Martín De la Fuente

OBJETIVOS

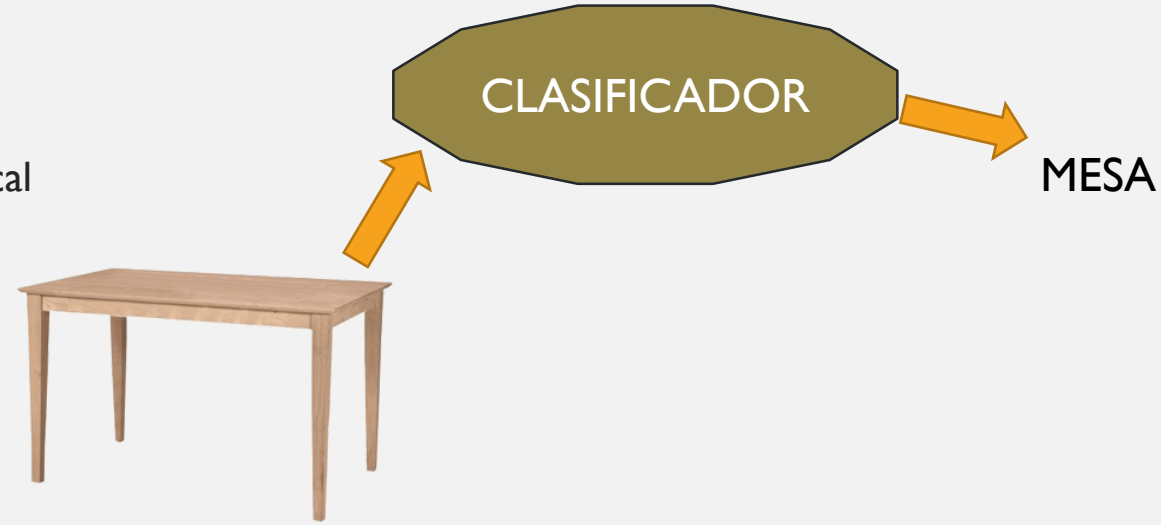
- Hacer resumen de los contenidos.
- Lograr entender bien qué estamos haciendo en la tarea.
- Manejar los términos que se usan en la tarea.
- Manejar conceptos para poder responder el informe.
- Avanzar en su tarea y resolver dudas.

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 generalmente la parte más difícil.
- 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 sobre profesionales en USA

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

Datos sobre profesionales en USA: identificar la clase

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

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 sobre profesionales en USA: selección de características

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

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 sobre profesionales en USA: *features y labels*

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

Matriz de características

Vector de clases

Datos sobre profesionales en USA: *features y labels*

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

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 sobre profesionales en USA: entrenamiento y pruebas

X_train

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica
52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?
34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico
25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States
32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States
38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States
43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States

y_train

<=50K
<=50K
<=50K
<=50K
<=50K
<=50K
<=50K
>50K
>50K
>50K
>50K
>50K
<=50K
<=50K
>50K
<=50K
<=50K
<=50K
<=50K
>50K

X_test

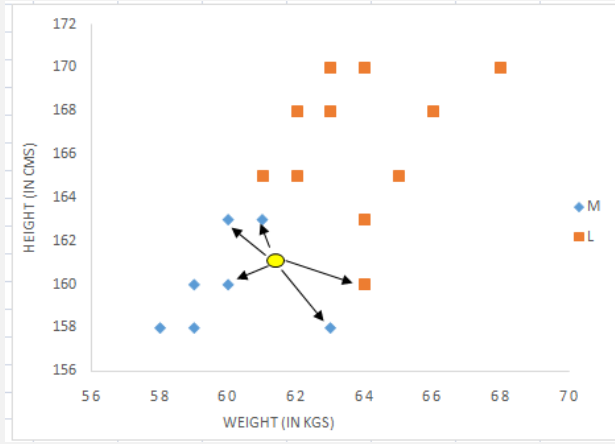
40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States
35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States
43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States
59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States

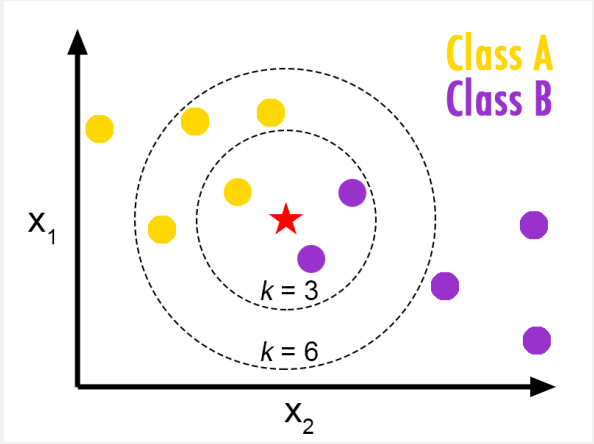
y_test

>50K
<=50K
<=50K
<=50K
<=50K

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.
- Es importante normalizar los datos.
- Parámetros importantes: K, pesos





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.

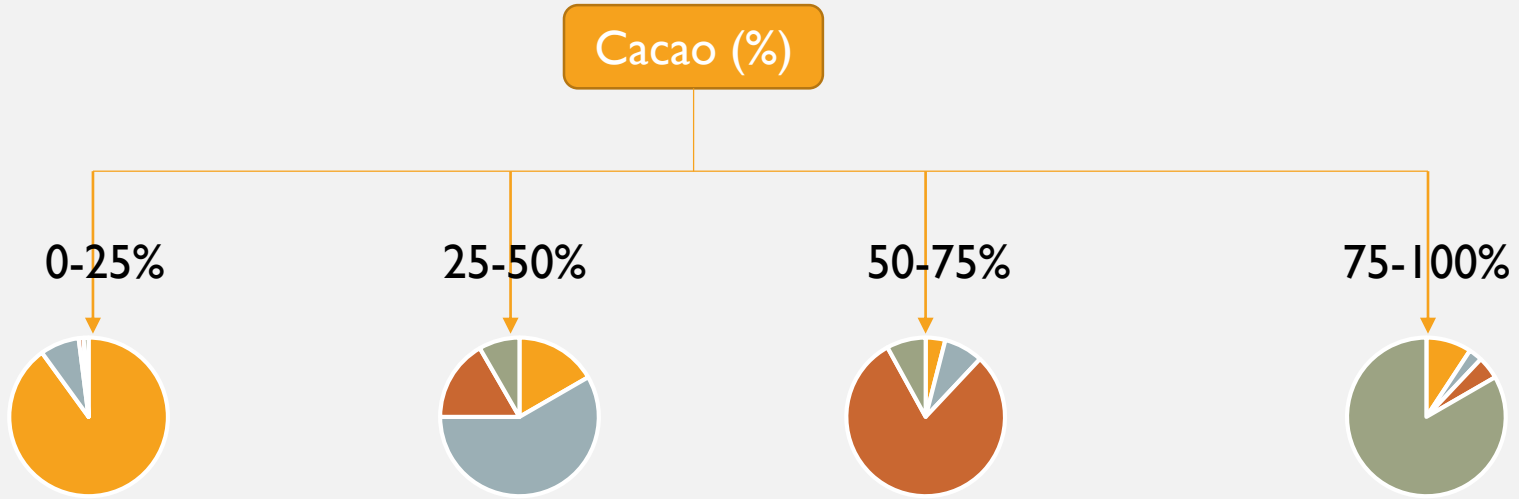
Cacao (%)	Leche (%)	Color	Precio (por 100gr)	Calidad
30	50	Café neutro	4312	Muy bueno
50	20	Café oscuro	4602	Regular
80	5	Café oscuro	8160	Muy bueno
20	40	Café neutro	2569	Malo
10	60	Café claro	1420	Malo
10	70	Café claro	1032	Bueno
30	20	Café neutro	4926	Bueno
60	10	Café oscuro	8741	Regular
55	5	Café 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

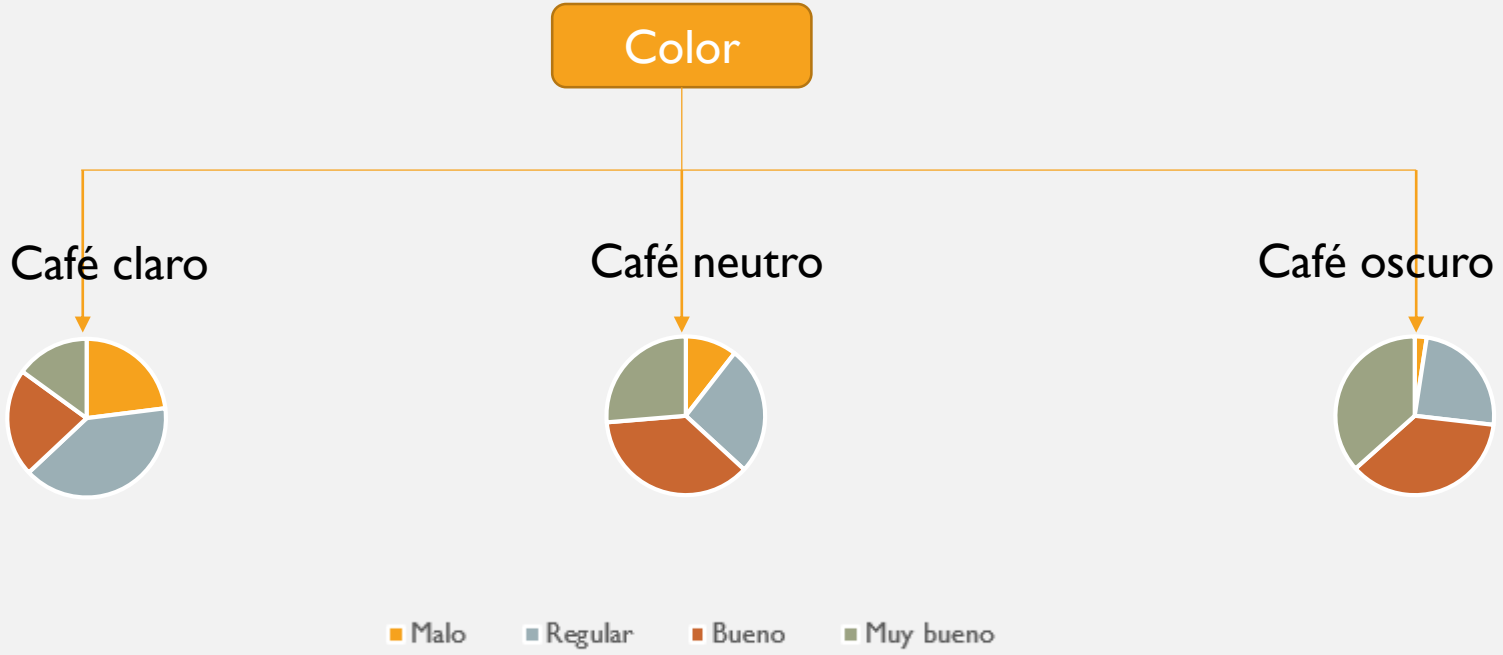
¿Qué atributo tiene mayor relación con la calidad?

¿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?



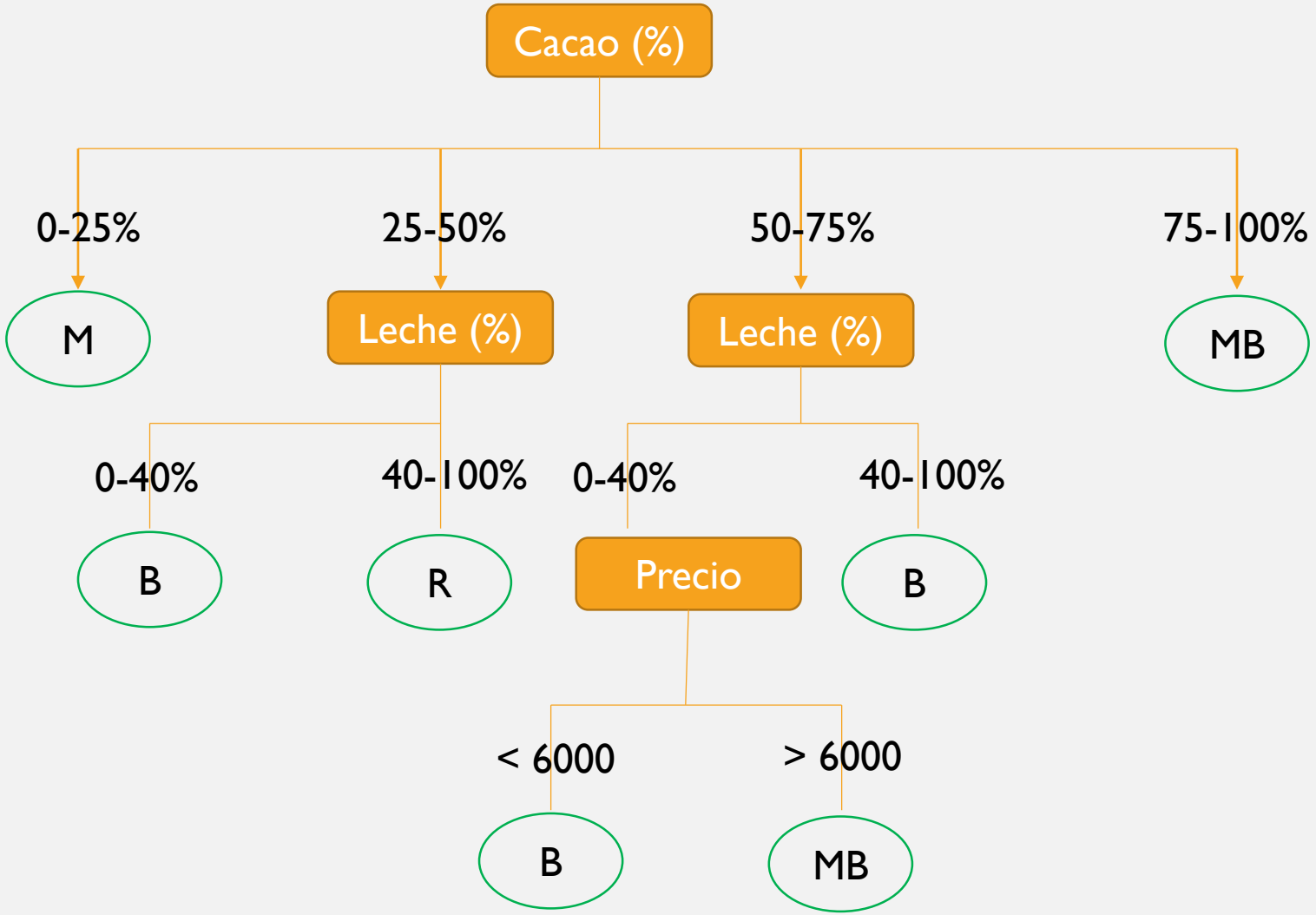


Cacao (%)	Leche (%)	Color	Precio (por 100gr)	Calidad
30	50	Café neutro	4312	Muy bueno
50	20	Café oscuro	4602	Regular
80	5	Café oscuro	8160	Muy bueno
20	40	Café neutro	2569	Malo
10	60	Café claro	1420	Malo
10	70	Café claro	1032	Bueno
30	20	Café neutro	4926	Bueno
60	10	Café oscuro	8741	Regular
55	5	Café 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

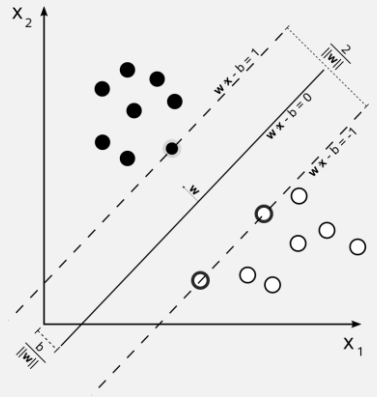
Situamos más cerca del nodo raíz a aquellos atributos que mejor separan las clases.

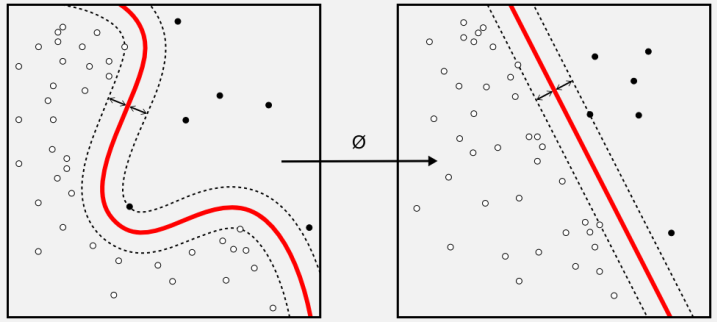
Mientras más profundo estamos en el árbol, menos ejemplos estaremos considerando. No es buena idea tomar la decisión cuando tenemos pocos ejemplos.



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**

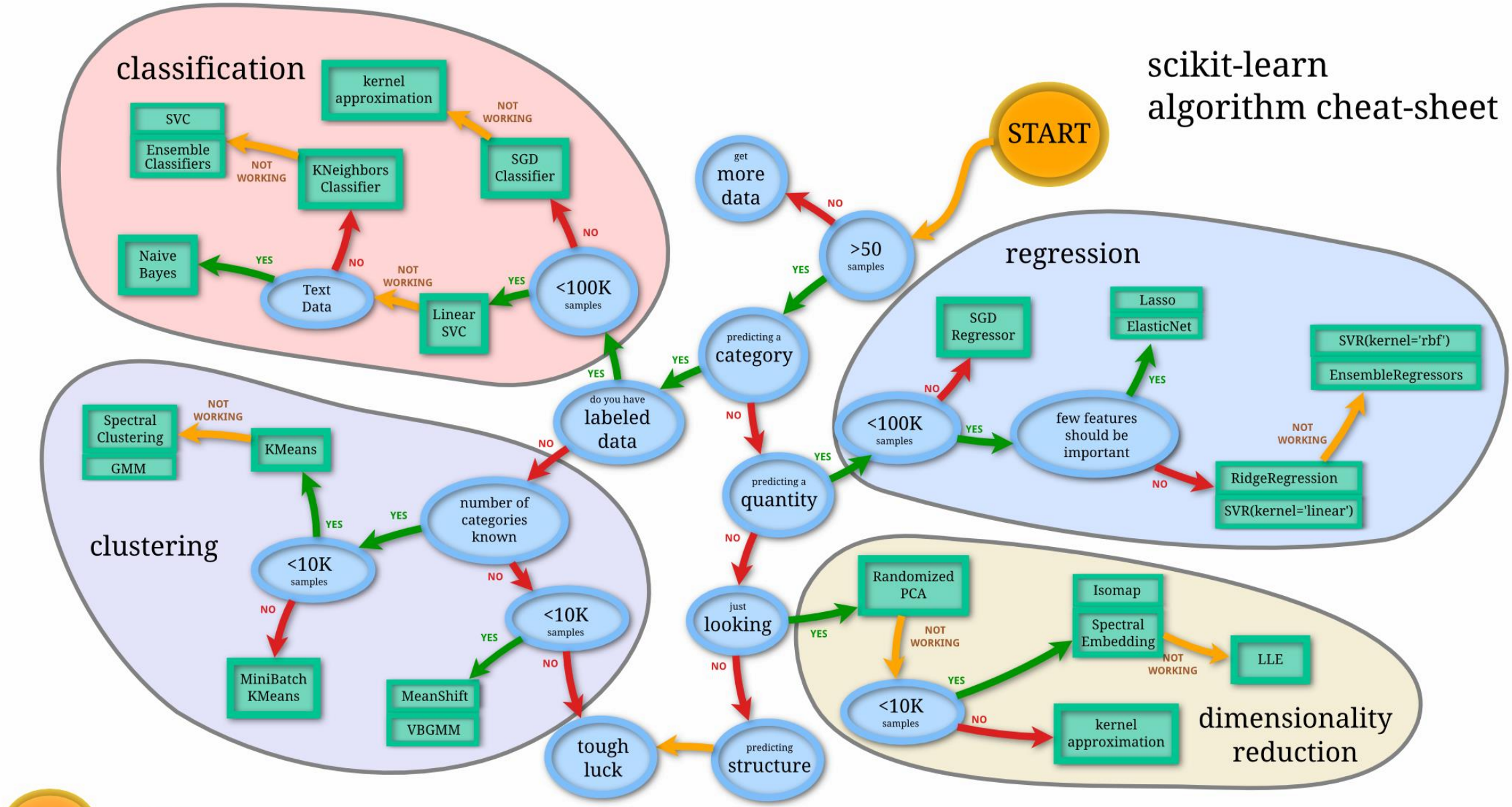
- Es importante que el set de pruebas tenga varios ejemplos de cada clase, y a su vez, la cantidad de ejemplos sea homogénea en todas las clases.
  - Se puede representar como valor numérico o como porcentaje.
- La MC permite ver qué clases son identificadas mejor y peor.
- La MC permite ver qué clases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

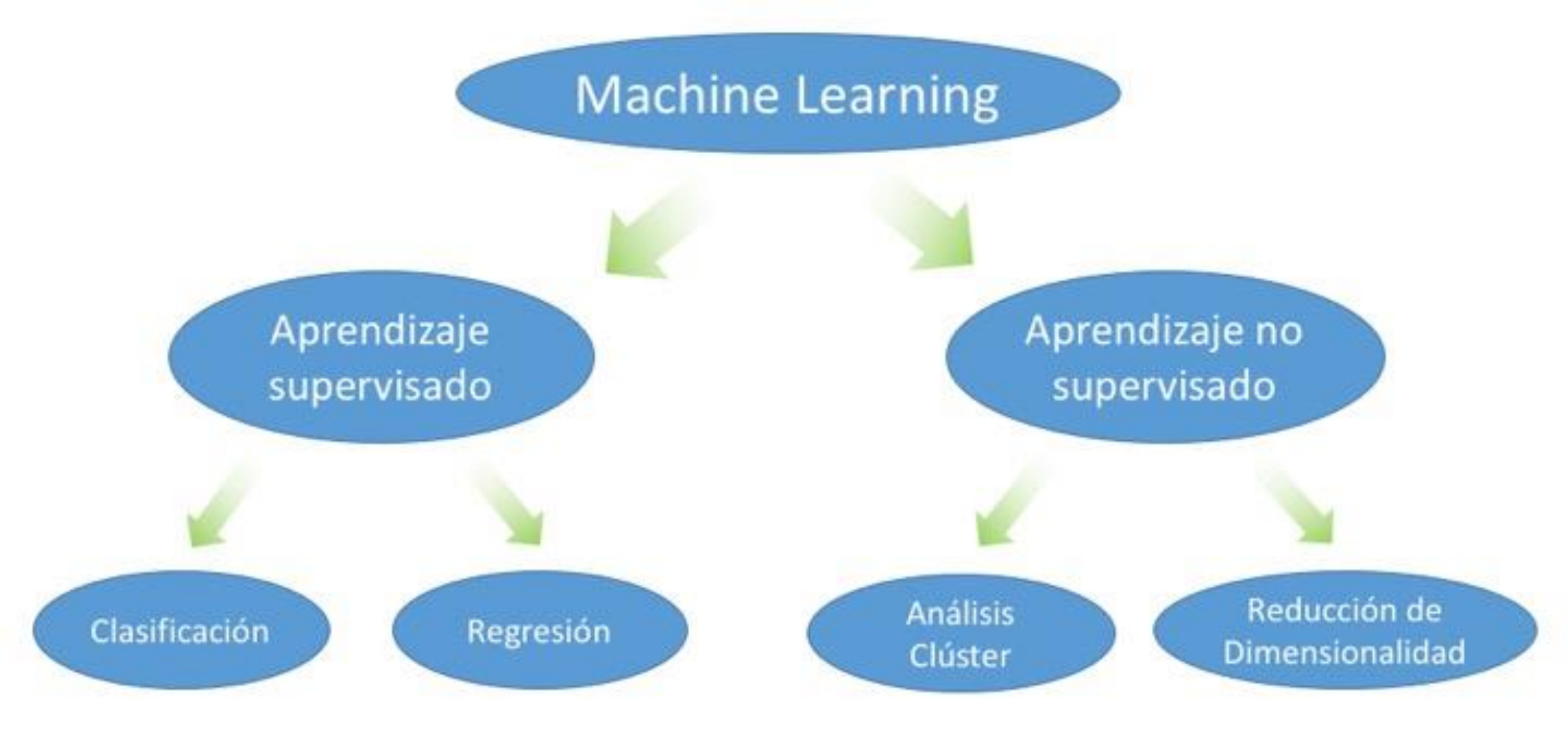
CLASIFICADORES

¿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

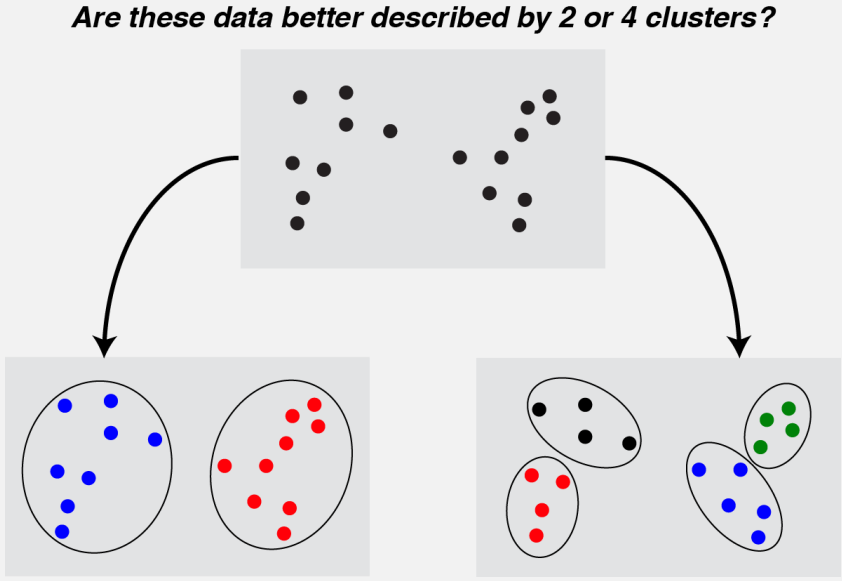
CLUSTERING

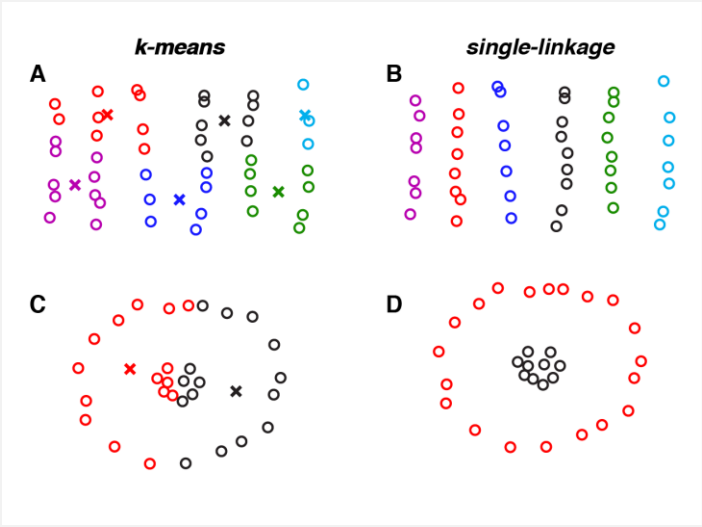
CLUSTERING

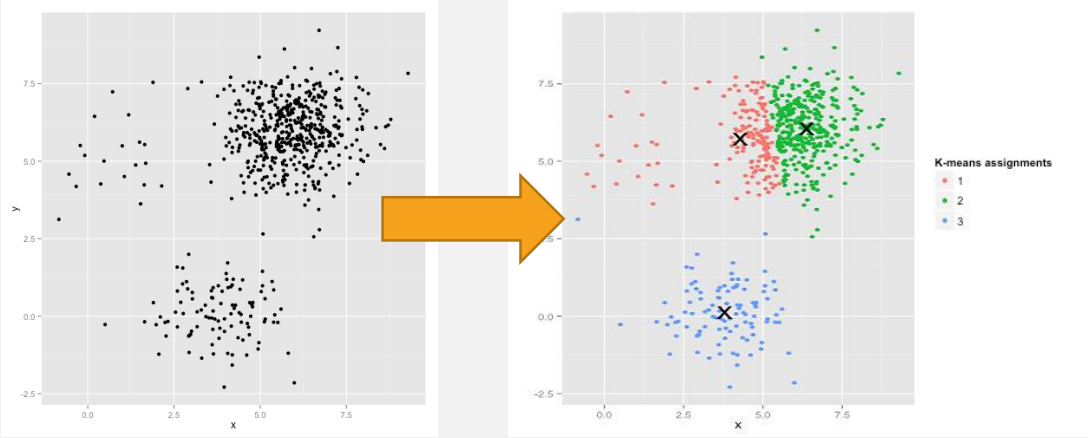
¿Para qué sirve?

- 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