Introducción al Aprendizaje Automático

Índice

Introducción

Importancia de los datos

Extracción de descriptores

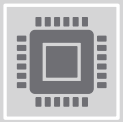
Tipos de aprendizaje

No supervisado

Supervisado

Índice

Aprendizaje Automático



El Aprendizaje Automático (Machine Learning) es una rama de la inteligencia artificial que ha transformado la forma en que abordamos tareas complejas.

El aprendizaje automático es la capacidad de las máquinas para aprender patrones a partir de datos y tomar decisiones sin ser programadas explícitamente.

Desde diagnósticos médicos hasta recomendaciones de películas, el aprendizaje automático está presente en diversas áreas.

PLN

PLN es un campo de la IA que se enfoca en la interacción entre las computadoras y el lenguaje humano. Permite a las máquinas comprender, interpretar y generar texto de manera similar a como lo hacen las personas.

Aprendizaje Automático y PLN

El Aprendizaje Automático es esencial en el PLN para tareas como:

- Traducción automática
- Análisis de sentimientos
- Resumen de texto
- Q/A
- Extracción de entidades

Clasificación de sentimientos

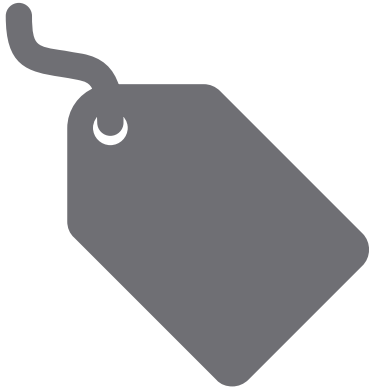
¿Cómo clasificamos el sentimiento de esta review?

Estimado proveedor de Amazon, la semana pasada pedí el libro "NLP with transformer" mediante un envío exprés y, desafortunadamente cuando abrí el paquete descubrí que no era el libro que esperaba. Además estaba arrugado y sucio. como consumidora, esto es un problema y solicito el reembolso de mi dinero. Espero su respuesta.
Un saludo.

Clasificación de sentimientos

3 posibles etiquetas:

- Si: más palabras positivas que negativas → **positivo**
- Si: más palabras negativas que positivas → **negativo**
- Si: mismo número de palabras negativas que positivas → **neutro**



Clasificación de textos

Problemas:

- Habría que crear muchas reglas
- Cada regla acabaría teniendo sus excepciones
- Negaciones
- Ironias

Clasificación de textos

Corpus de datos

¿Cuando se supone que aprendemos nosotros?

Obtenemos conocimiento mediante el estudio, la experiencia o al ser enseñado

Fijamos algo en la memoria

Nos damos cuenta de algo a través de cierta información o a partir de la observación

“El aprendizaje automático es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a los ordenadores aprender sin ser programados de manera explícita” [A. Samuel, 1959]

¿Cuando aprende una máquina?

Más ejemplos, mejor desarrollo de la tarea

Prueba - error, se adapta

Entrenamiento

Aproximación resolución problemas

Aproximación tradicional programación:

- Desarrollar programa que resuelva una tarea con REGLAS

Limitaciones:

- Sólo resuelven problemas ya previstos

- Un sistema se considera inteligente si es capaz de observar su entorno y aprender de él
- Inteligencia reside en adaptarse, tener capacidad de integrar nuevo conocimiento, resolver nuevos problemas, y aprender de los errores

Aproximación resolución problemas

Aproximación aprendizaje automático:

- Buscamos dar a los programas la capacidad de adaptarse sin tener que ser reprogramados, PATRONES

- No imita el aprendizaje humano

Aproximación resolución problemas

Aproximación aprendizaje automático:

1. Crear algoritmo con una serie de parámetros
2. Obtener conjunto de ejemplos de entrenamiento que especifica parcialmente comportamiento deseado del sistema
3. Algoritmo toma ejemplos y fija los parámetros de modo que es capaz de producir de manera aproximada el comportamiento del sistema

Aproximación resolución problemas

Magia aprendizaje automático:

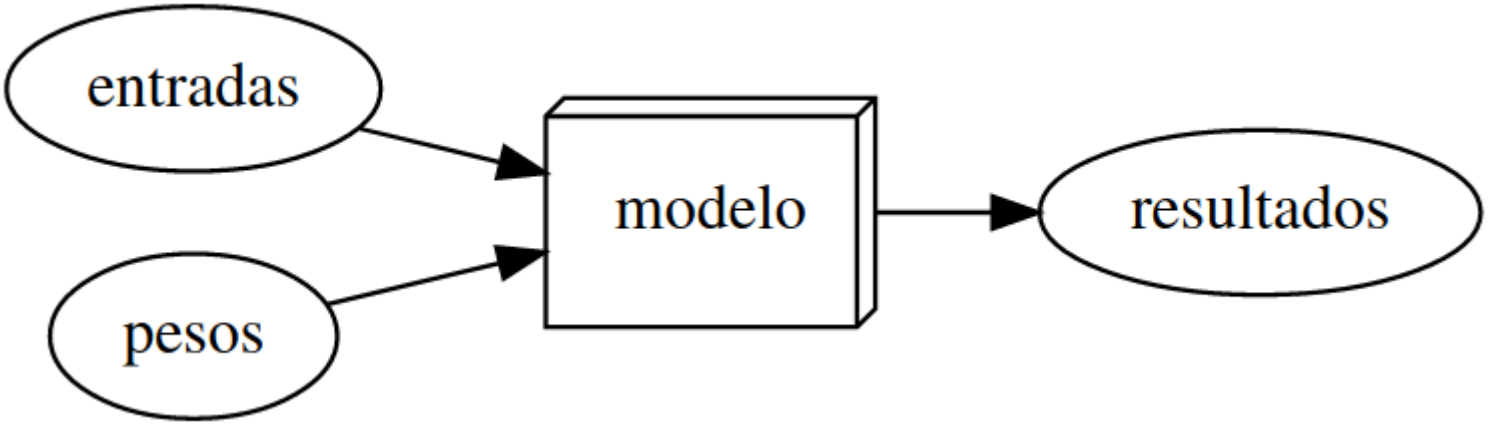
- Si se realiza de manera correcta:
  - Algoritmo hace **buenas predicciones** para conjunto entrenamiento
  - Y además **generaliza** bien

Aprendizaje automático vs Programación

Programación tradicional

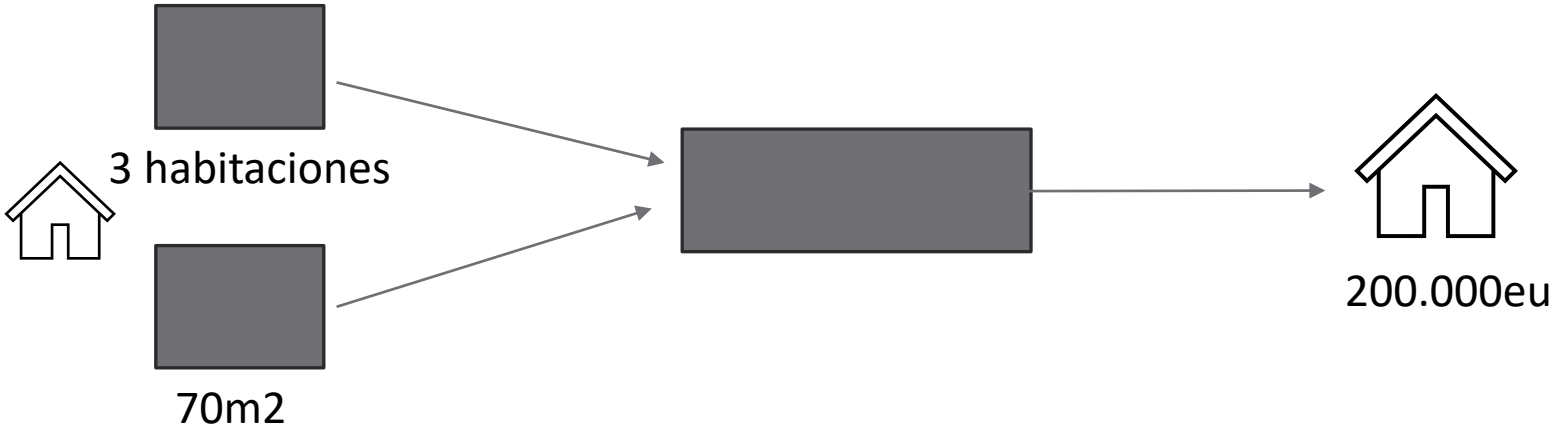
Aprendizaje automático vs Programación

Aprendizaje automático



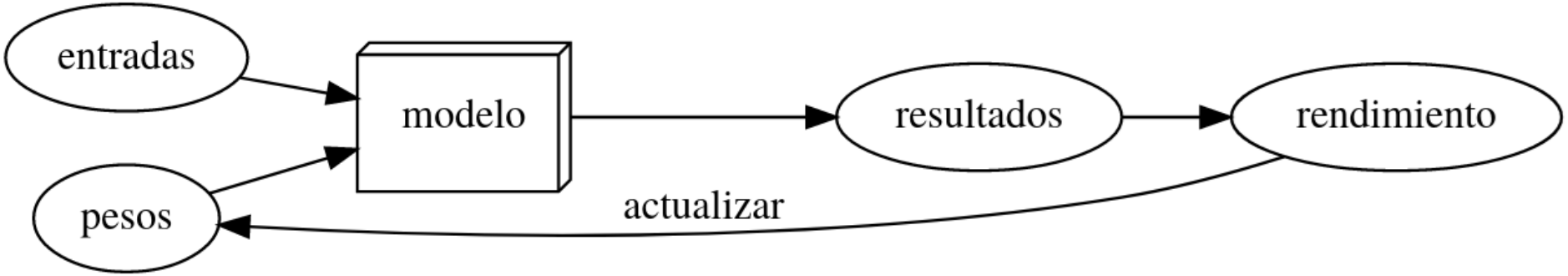
Aprendizaje automático vs Programación

Aprendizaje automático



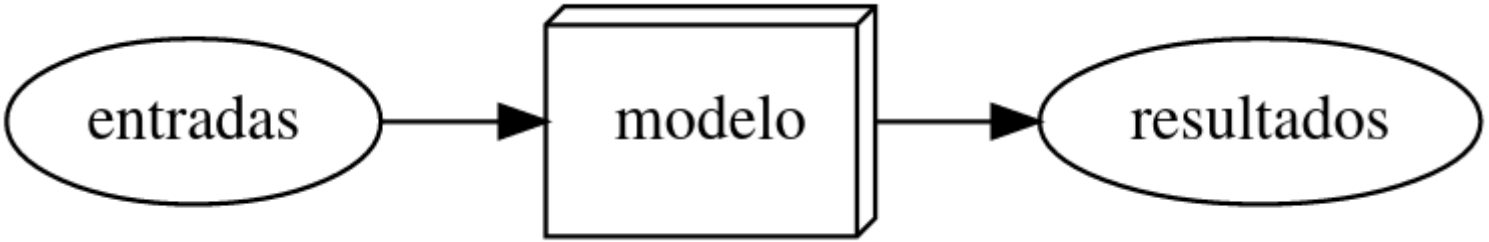
Aprendizaje automático vs Programación

Aprendizaje automático



Aprendizaje automático vs Programación

Aprendizaje automático



Aplicaciones

Tareas difíciles de programar:

- Adquisición de conocimiento, reconocimiento de caras, de voz, …

Aplicaciones auto adaptables

- Sistemas de recomendación, detector de spam, interfaces inteligentes, …

Terminos Aprendizaje Automático

- Lo que queremos aprender → descripción del **concepto**
- Información que se transmite al sistema → **instancias**
  Instancia: ejemplo individual e independiente del concepto que se quiere aprender
  - Cada instancia se caracteriza por los **descriptores** (o atributos) que miden distintos aspectos de cada instancia

Índice de contenidos

Importancia de los datos

"LOS DATOS SON EL NUEVO PETRÓLEO, Y LA IA LA NUEVA ELECTRICIDAD" [ANDREW NG]

Captura de datos

Es el primer paso en cualquier aplicación de aprendizaje automático, puede ser complicado

Capturar instancias que formen nuestro banco de datos (*dataset)*

Avance de aprendizaje automático gracias a grandes datasets → permiten generalizar

Bancos de datos representativos

QUEREMOS CREAR SISTEMA CAPAZ DE PREDECIR MARCA Y MODELO DE COCHES QUE VAN POR LA AUTOPISTA

Bancos de datos representativos

Cogemos imágenes de la web

Bancos de datos representativos

Construimos sistema con alta tasa de acierto en imágenes de coches, pero al poner el sistema en funcionamiento en la autopista no funciona ¿por qué?

Bancos de datos representativos

Construimos sistema con alta tasa de acierto en imágenes de coches, pero al poner el sistema en funcionamiento en la autopista no funciona ¿por qué?

Banco de datos usado para preparar sistema no es representativo

Índice de contenidos

Descriptores (features)

Instancia→vector de descriptores (feature vectors)

- Para cada instancia toman distintos valores
- Para todas las instancias se estudian las mismas propiedades
- Ejemplo: descripción casa:
  - nº habitaciones, nº baños, localización
  - Todas las casas descritas con esos tres descriptores, no es posible usar distintos descriptores para cada casa

Extracción de descriptores

Depende del dominio y de la información que se pueda adquirir

Puede ser un proceso costoso

Descriptores

- Natural trabajar con bases de datos estructuradas
- ¿Qué ocurre con imágenes, vídeos, texto, audio, …?
  - Datos no estructurados
  - Usar **embeddings**
  - Datos no estructurados en vectores numéricos


INPUT DATA
Garbage in
Garbage OUT
OUTPUT DATA
ALGORITMO
PERFECTO

Datasets

Expectativa

Realidad

PHIADELPHIA	PHILADELPOHIA
PHIALDELPHIA	PHILADELPPHIA
PHIDELPHIA	PHILADEPHA
PHIELADELPHIA	PHILADEPHIA
PHIILADELPHIA	PHILADEPHILA
PHILA	PHILADEPLHIA
PHILA.	PHILADERLPHIA
PHILAD	PHILADLELPHIA
PHILADALPHIA	PHILADLEPHIA
PHILADEDLPHIA	PHILADLPHIA
PHILADELAPHIA	PHILADPHIA
PHILADELHIA	PHILADRLPHIA
PHILADELHPIA	PHILAELPHIA
PHILADELLPHIA	PHILDADELPHIA
PHILADELOHIA	PHILDADLPHIA
PHILADELPH	PHILDAELPHIA
PHILADELPHA	PHILDELPHIA
PHILADELPHAI	PHILDEPPHIA
PHILADELPHI	PHILIADELPHIA
PHILADELPHIA	PHILIDELPHIA
PHILADELPHIA PA	PHILLA
PHILADELPHIA,	PHILLADELPHIA
PHILADELPHIA, PA	PHILLY
PHILADELPHIA`	PHILOADELPHIA
PHILADELPHIAP	PHLADELPHIA
PHILADELPHIAPHIA	PHOLADELPHIA
PHILADELPHILA	PHPILADELPHIA
PHILADELPHIOA	PIHLADELPHIA
PHILADELPIA	

‘Dirty data’ problems

Inaccurate	Incomplete	Inconsistent	Incompatible
Data stored as wrong type	Uncategorised	Inconsistency in naming of entities	Wrong shape
Misentered data	Missing data	Mixed data	‘Dirty’ characters (e.g. unescaped HTML)
Duplicate data			
Abbreviation and symbols			

Dirty

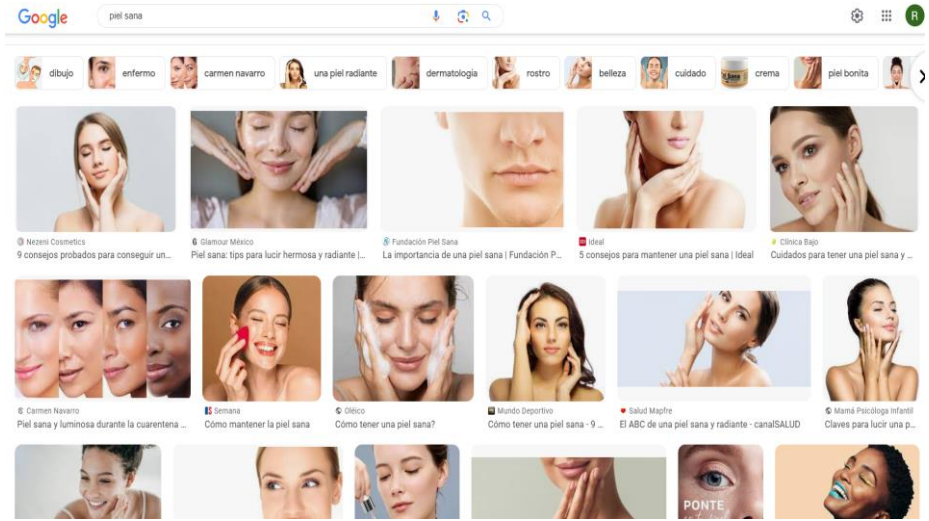
Ship Mode	First Class			Same Day	
Segment	Consumer	Corporate	Home Office	Consumer	Corporate
Order Date					
14-Mar-13					
16-Dec-13					
02-Jun-13					
21-Oct-13					
27-Aug-13					
28-Nov-13					
31-Mar-13					
21-Nov-13					
01-Nov-13					
05-Apr-13					
05-Jul-13		242.546			
15-Jan-13	149.95				
02-Dec-13					
19-Mar-13		590.762			
27-Jun-13					
06-Jan-13		12.78			
14-May-13					
12-Dec-13					
29-Apr-13					

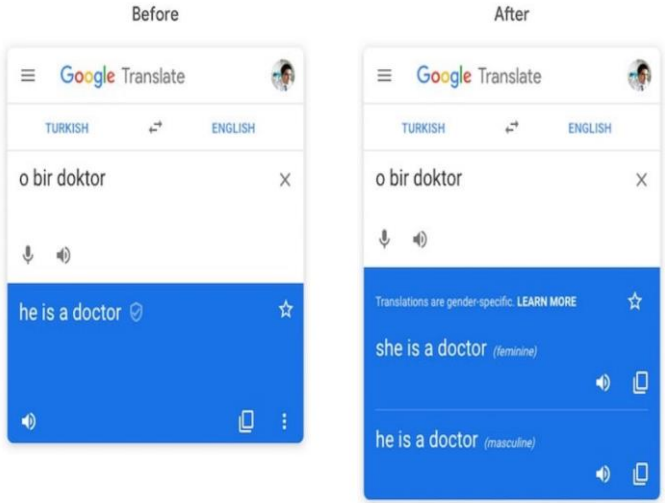
Clean

Ship Mode	Segment	Order Date	Sales
First Class	Consumer	15-Jan-13	149.95
First Class	Consumer	15-Aug-13	243.6
First Class	Consumer	24-Dec-13	9.568
First Class	Consumer	07-Apr-13	8.96
First Class	Consumer	19-May-13	34.2
First Class	Consumer	05-Sep-13	31.984
First Class	Consumer	12-Aug-13	286.65
First Class	Consumer	05-Jul-13	514.03
First Class	Consumer	30-Apr-13	1000.95
First Class	Consumer	23-Mar-13	9.912
First Class	Consumer	30-Dec-13	39.128
First Class	Consumer	18-Apr-14	106.5
First Class	Consumer	21-Nov-14	18.176
First Class	Consumer	23-Dec-14	194.32
First Class	Consumer	23-Mar-14	59.48
First Class	Consumer	30-Oct-14	182.91
First Class	Consumer	16-Apr-14	2298.9
First Class	Consumer	02-Nov-14	197.72
First Class	Consumer	30-Nov-14	440.144
First Class	Consumer	22-Nov-14	32.985
First Class	Consumer	11-Dec-14	196.62

Errores en los datos

Sesgos en los datos





Índice de contenidos

Aprendizaje no supervisado

El aprendizaje no supervisado es una rama del aprendizaje automático en la que el algoritmo se entrena en datos sin etiquetas ni categorías predefinidas.

Ejemplo: Agrupación de textos en categorias basados en su contenido sin conocer esas las categorías.

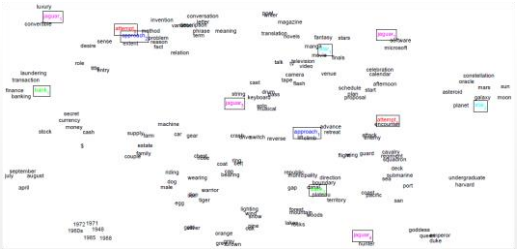
Aprendizaje no supervisado

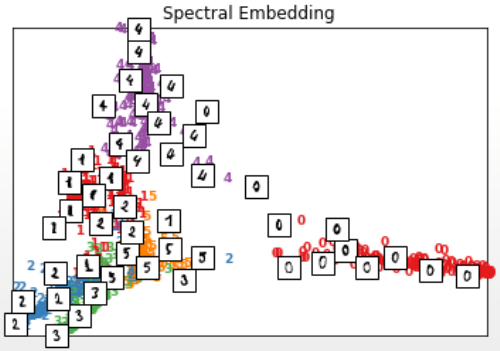
Entrada:

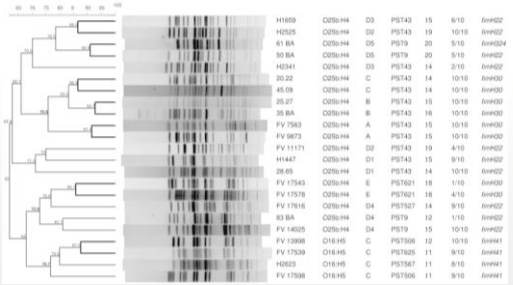
- Dataset con los vectores de descriptores

Objetivo:

- Encontrar alguna estructura, patrón o relación entre las instancias del dataset, pero sin conocer dicha estructura de antemano

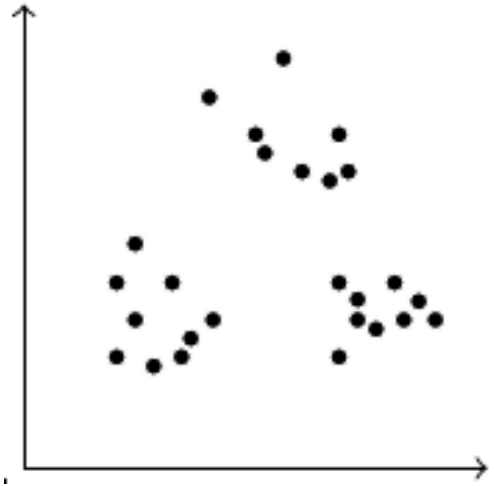


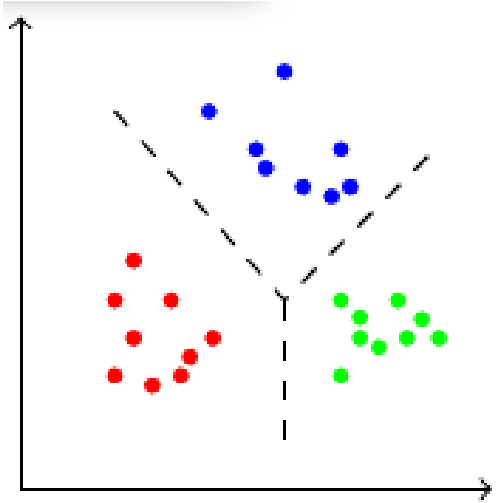

Spectral Embedding

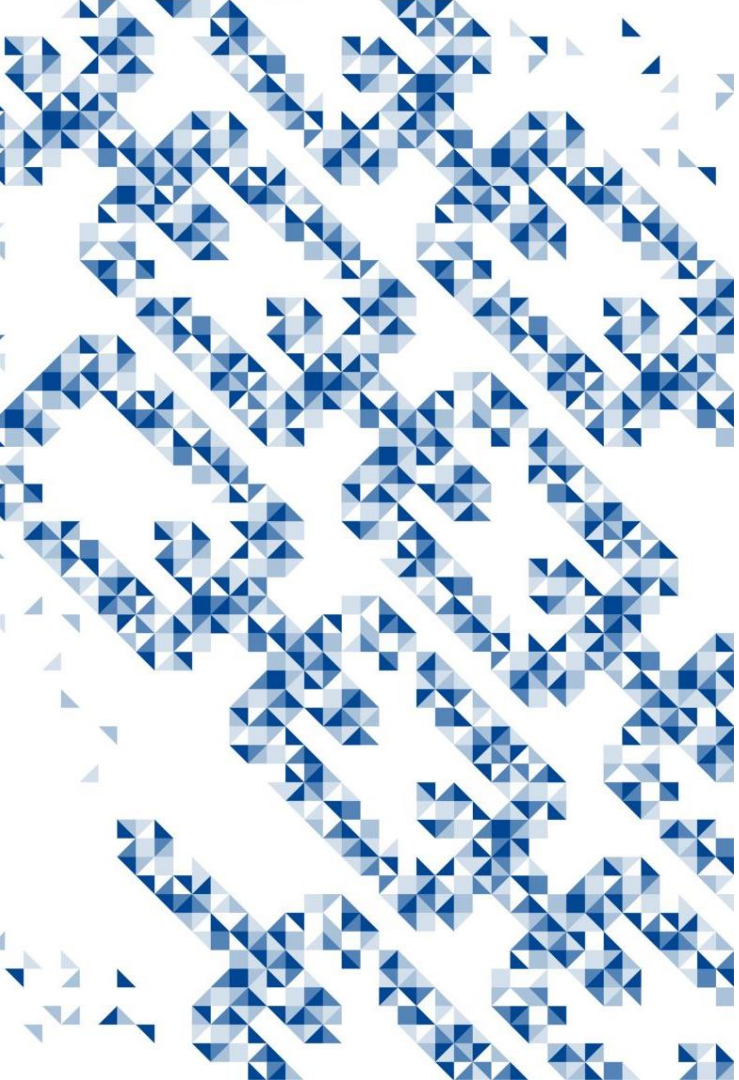


Clustering (o agrupamiento)

Técnica de aprendizaje no supervisado que busca agrupar datos similares en conjuntos llamados clústeres



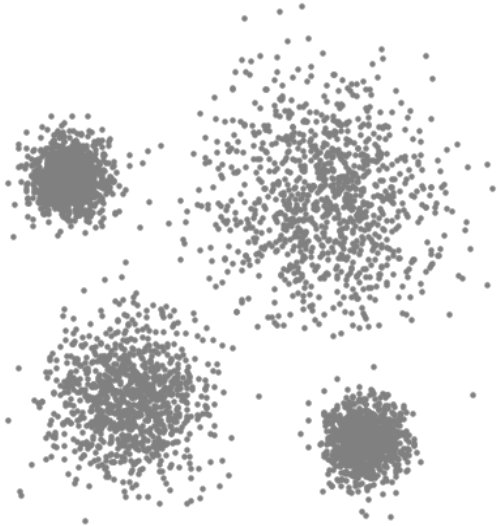


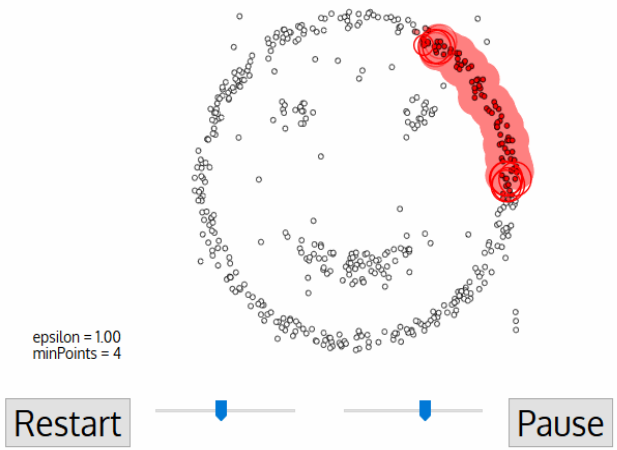


Clustering

- ¿Y si aumentamos el número de datos?
- ¿Y si aumentamos las dimensiones?

K-means, Mean shift, DBSCAN





Reducción de la dimensionalidad

Instancias:

- Puntos en un espacio N-dimensional
- N es el número de descriptores

Problemas con N grande:

- Información redundante
- Visualización de datos

- Reducen complejidad de los modelos
- Resuelven problema del sobreajuste

Reducción de la dimensionalidad

- PCA: El Análisis de Componentes Principales (PCA) es una técnica común de reducción de dimensionalidad que identifica las direcciones en las que los datos varían más y las proyecta en un espacio de menor dimensión.

Aprendizaje supervisado

El aprendizaje supervisado es un enfoque del aprendizaje automático en el que el modelo se entrena en un conjunto de datos etiquetados, es decir, datos que tienen una respuesta o salida conocida.

Ejemplo: Clasificación de correos electrónicos como spam o no spam en función de ejemplos previamente etiquetados

Aprendizaje supervisado

Objetivo: a partir de una entrada aprender una salida

Entrada	Salida	Aplicación
Email	¿Spam?	Filtro anti-spam
Audio	Texto transcrito	Reconocimiento del habla
Texto en inglés	Texto en español	Traducción automática
Anuncio e información usuario	¿Hizo click?	Publicidad online
Imagen de un teléfono	¿Defectuoso?	Inspección de errores

Aprendizaje supervisado

Objetivo:

- A partir de un conjunto de entrenamiento etiquetado
- Construir un modelo capaz de predecir la etiqueta de nuevas instancias

Aprendizaje supervisado vs no supervisado

Dataset aprendizaje supervisado:

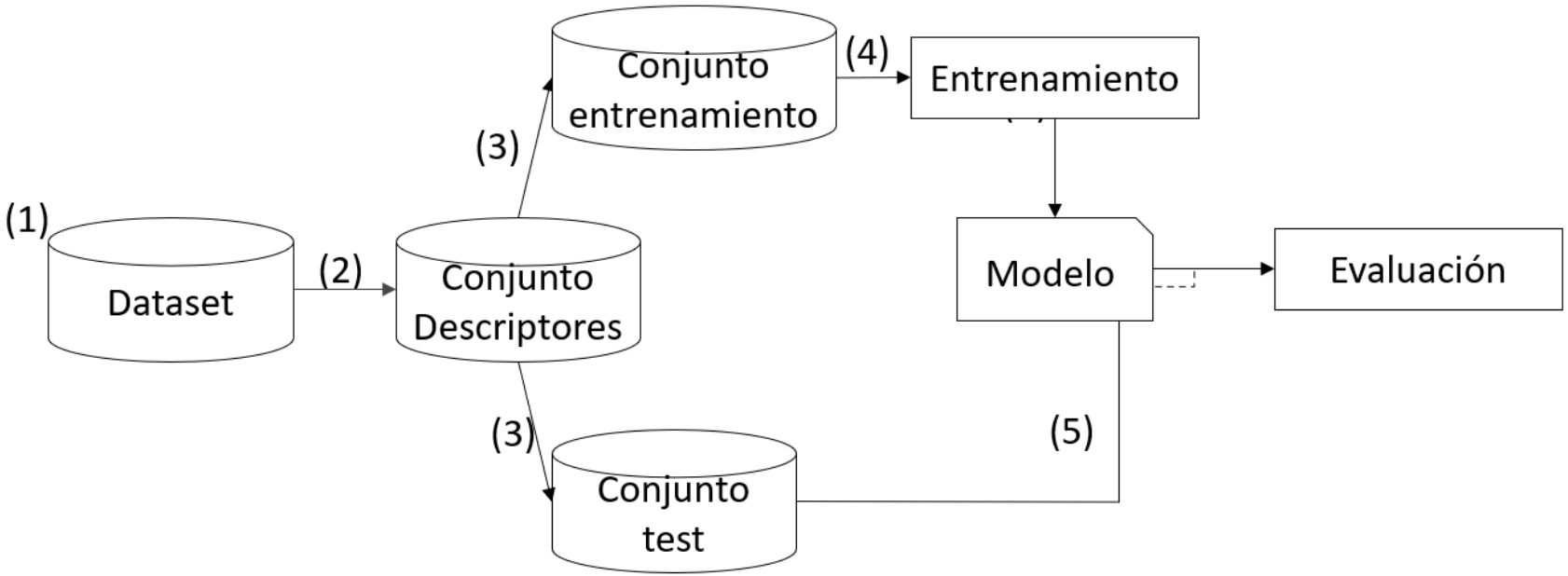
- Datos de cada instancia
- Etiqueta asociada a cada instancia:
  - Spam/No spam
  - Precio de la vivienda
  - Dígito manuscrito
  - …

Flujo de trabajo

Proceso de aprendizaje supervisado:

1. Estructurar dataset inicial
2. Extraer descriptores
3. Partir dataset en dos (o tres) partes
4. Entrenar/construir un modelo de predicción
5. Evaluar el modelo

Flujo de trabajo



Recolectando y estructurando

- Además de instancias necesitamos sus etiquetas
- Número de instancias por categoría:
  - Debería ser uniforme
  - A mas ejemplos, mejor

¿Cómo etiquetar los datos?

- De manera manual
- Descargando datasets ya anotados
- De manera automática observando comportamientos

Partición del dataset

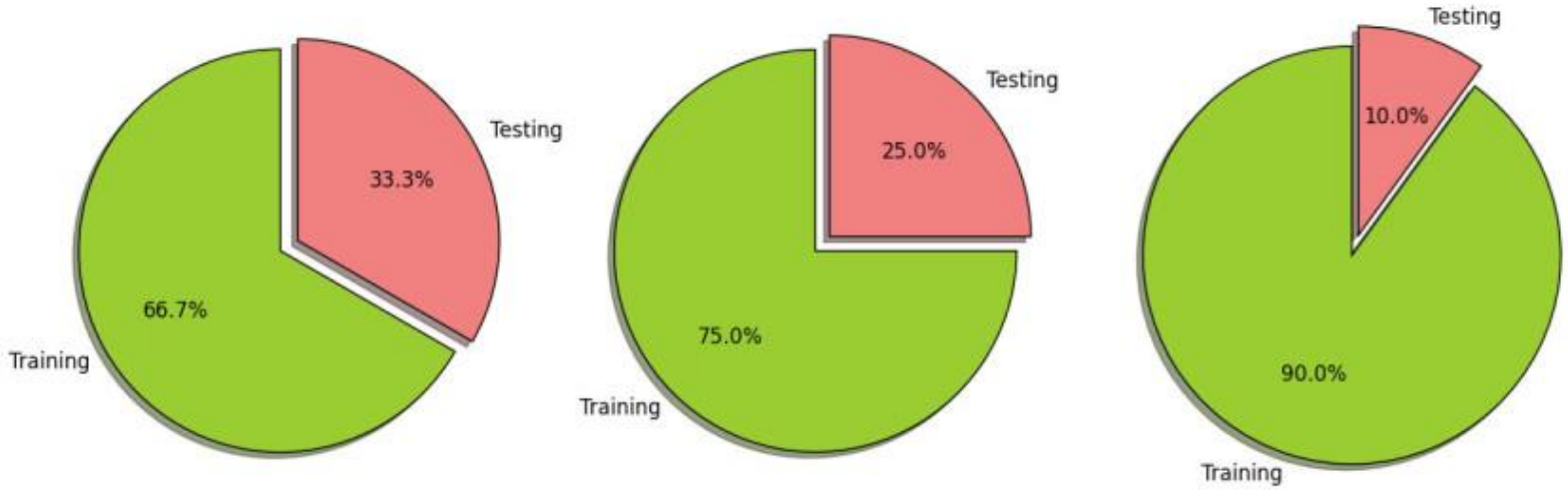
Dataset inicial se divide en dos partes:

- Conjunto de entrenamiento
- Conjunto de test

Importante:

- Conjuntos deben ser independientes

Partición del dataset



Partición conjunto entrenamiento

Modelos de aprendizaje supervisado dependen de parámetros llamados hiperparámetros:

- Necesario ajustarlos
- Probar varios para ver cual produce mejores resultados
- No se puede usar conjunto de test
- Conjunto de entrenamiento se parte en dos: entrenamiento (80-90%) y validación (20-10%)

Entrenando/construyendo modelos

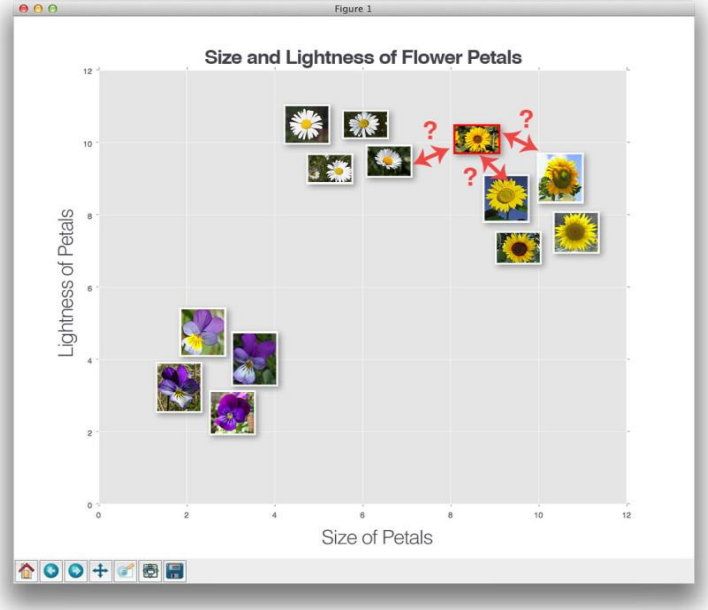
Existen diversos algoritmos:

- KNN
- Árboles de decisión
- SVMs
- Regresión lineal
- Regresión logística
- Redes neuronales
- ...

KNN

- Algoritmo de los k vecinos más cercanos
- K-Nearest Neighbor o KNN
- Algoritmo de clasificación más sencillo
  - En realidad no aprende nada, no se entrena
  - Basado en distancia entre vectores de descriptores
  - Estilo k-means pero con etiquetas
  - Clasifica nuevas instancias encontrando clase más común entre los k ejemplos más cercanos
  - “Dime con quién andas y te diré quién eres”

KNN



Hiperparámetro: los vecinos más cercanos, ej:3

Valor de k: siempre impar

Regresión lineal

Idea intuitiva: añadir los efectos de cada descriptor para obtener el valor predicho

Regresión lineal

Cesta de la compra:

- 2.5kg de patatas, 1kg de zanahorias, 2 bricks de leche
- Patatas a 2€ el kg, Zanahorias a 4€ el kg, cada brick de leche 1€
- ¿Cuál es el precio total?
  - 2.5*2+1*4+2*1 = 11€

Regresión lineal

- Cantidad de patatas, zanahorias y leche son los datos de entrada (descriptores)
- Precios de diferentes productos son los coeficientes o pesos
- El coste de la compra es la salida que depende de manera lineal del precio de cada producto y la cantidad

Regresión lineal

Cesta de la compra con 1kg de ternera, 2kg de zanahorias y 1 botella de vino, vale 35€ ¿cuánto cuesta cada cosa?

Con un único ejemplo no lo podemos resolver, pero si tenemos varias cestas sí que se puede abordar el problema

Regresión lineal

A partir de:

- Datos de entrada y salida
- Aprende los pesos para predecir salida a partir de nuevos datos de entrada

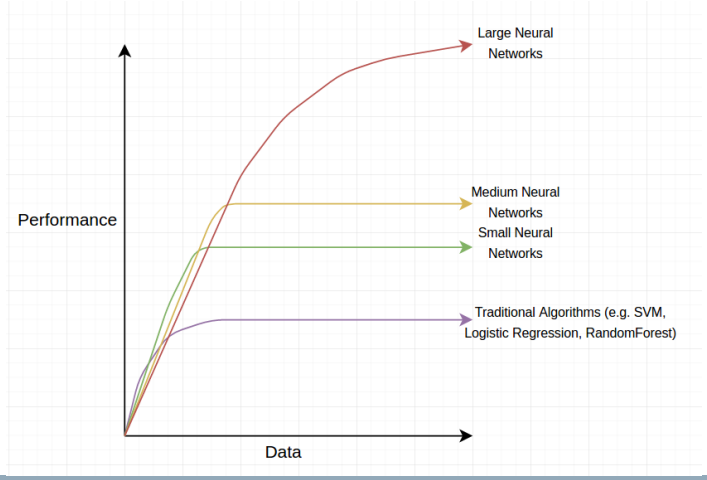
Regresión lineal

Mundo real:

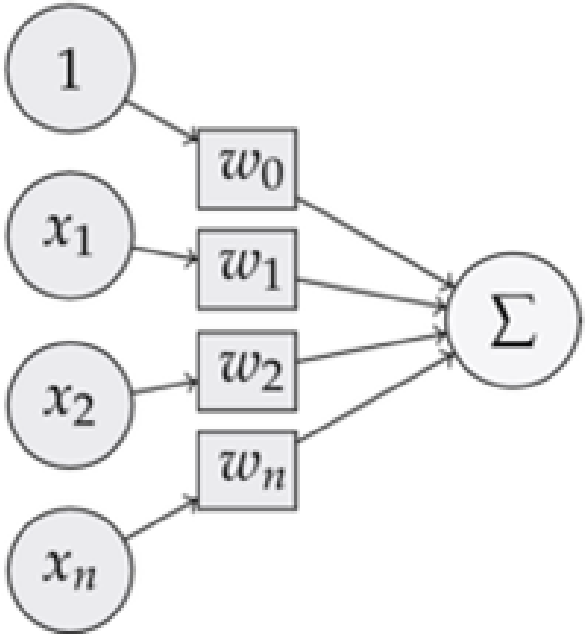
- No siempre se tiene toda la información que afecta a una salida
- Ciertos factores introducen ruido
- Solo es posible estimar salida hasta cierto punto

Redes neuronales artificiales

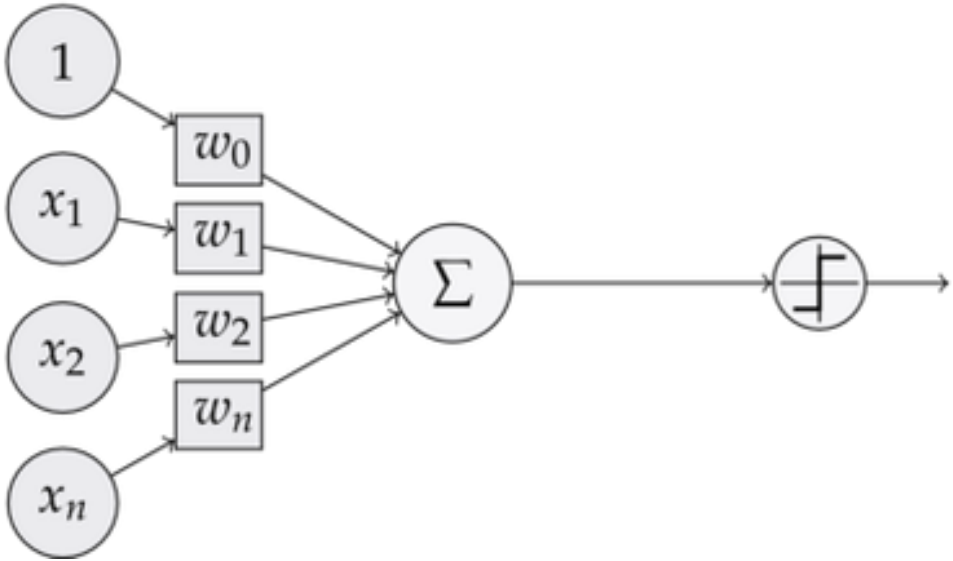
- Algoritmo creado en los años 50
- Ha ganado importancia recientemente gracias al deep learning



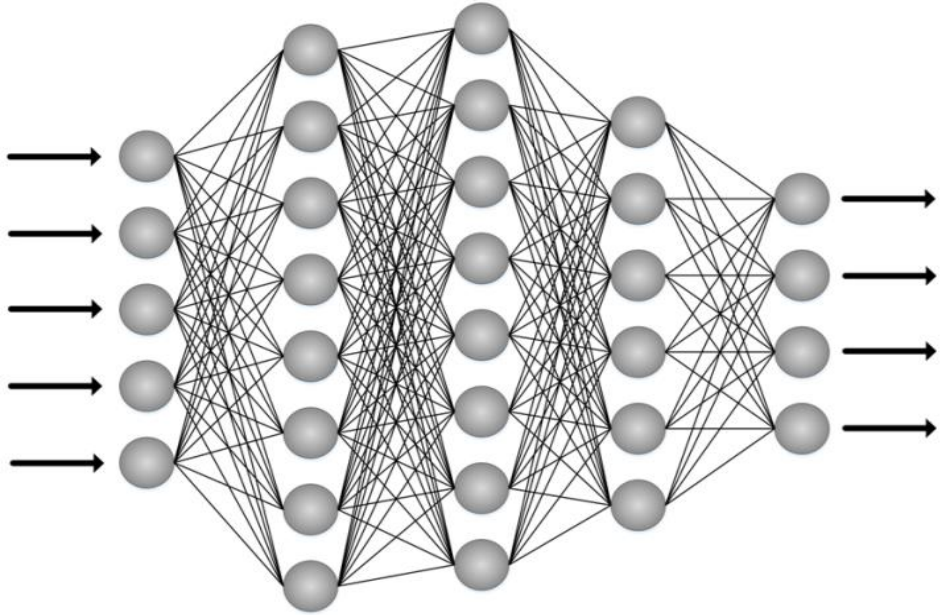
El perceptrón simple



Funciones de activación



Múltiples capas



Evaluando los modelos

Distintos modelos para clasificar:

- ¿Cuál elegimos?
- Solución → Evaluación:
  - Probar distintos modelos
  - Probar distintos hiperparámetros
  - Ver cuál funciona mejor en problema concreto

Métricas evaluación: caso binario

Rendimiento clasificador:

- Para cada instancia predecir su clase:
  - Acierta → Éxito
  - Falla → Error
- Ratio de error:
  - Proporción de errores sobre total
- Ratio éxito:
  - Proporción éxitos sobre total
- Ratios no tienen en cuenta coste:
  - Persona con enfermedad → no se le detecta
  - Persona sin enfermedad → se le detecta

Métricas evaluación: caso binario

		Clase predicha	
		sí	no
Clase real	sí	True positive	False negative
	no	False positive	True negative

- Verdaderos positivos → número de resultados predichos como sí que son sí
- Verdaderos negativos → número de resultados predichos como no que son no
- Falsos positivos → número de resultados predichos como sí cuando en realidad son no
- Falsos negativos → número de resultados predichos como no cuando en realidad son sí

Métricas evaluación: caso binario

Ratios a partir de valores anteriores:

- Ratio verdaderos positivos (sensitivity, recall, ...): TP/(TP+FN)
- Ratio verdaderos negativos (specificity,...): TN/(FP+TN)
- Precisión: TP/(TP+FP)
- F-measure: 2TP/(2TP+FP+FN)
- Accuracy: (TP + TN)/(TP+FP+FN+TN)

Métricas evaluación: caso múltiple

Se utiliza la matriz de confusión

		Clase predicha			
		a	b	c	Total
Clase real	a	88	10	2	100
	b	14	40	6	60
	c	18	10	12	40
	Total	120	60	20	

Nos interesa que esta matriz sea diagonal

Proceso de evaluación

Resumen:

- Conjunto entrenamiento → se usa para crear uno o varios clasificadores
- Conjunto de validación → se utiliza para optimizar hiperparámetros de los clasificadores
- Conjunto test → se usa para calcular ratio de error del clasificador

División en tres partes se conoce como método de holdout

Estratificación

Problema: muestra no representativa:

- Cada clase del dataset debería estar representada en proporción correcta tanto en conjunto de test como en el de entrenamiento
- Si conjunto entrenamiento no tiene instancias de una clase, clasificador no funcionará bien

Solución:

- Asegurarnos de que el muestreo aleatorio garantiza que cada clase está representada de manera adecuada tanto en conjunto entrenamiento como en el de test
- A este procedimiento se le conoce como estratificación

LAB