GOBIERNO DE ESPAÑA
MINISTERIO DE INDUSTRIA, COMERCIO Y TURISMO
EOI Escuela de organización industrial
UE23 PRESIDENCIA ESPAÑOLA CONSEJO DE LA UNIÓN EUROPEA
Unión Europea
Fondo Social Europeo
Iniciativa de Empleo Juvenil
El FSE invierte en tu futuro
SISTEMA NACIONAL DE GARANTÍA JUVENIL | SEPE

INTELIGENCIA ARTIFICIAL Y MACHINE LEARNING

Introducción al Machine Learning

Contenido

Introducción y definiciones ML

1. Presentación
2. Definición y usos
3. Tipos, descripción y ejemplos
4. Dataset: entrenamiento, validación y test
5. Métricas, bajoajuste y sobreajuste
6. Validación cruzada y validación cronológica
7. Algoritmos de optimización
8. Lenguajes de programación y plataformas

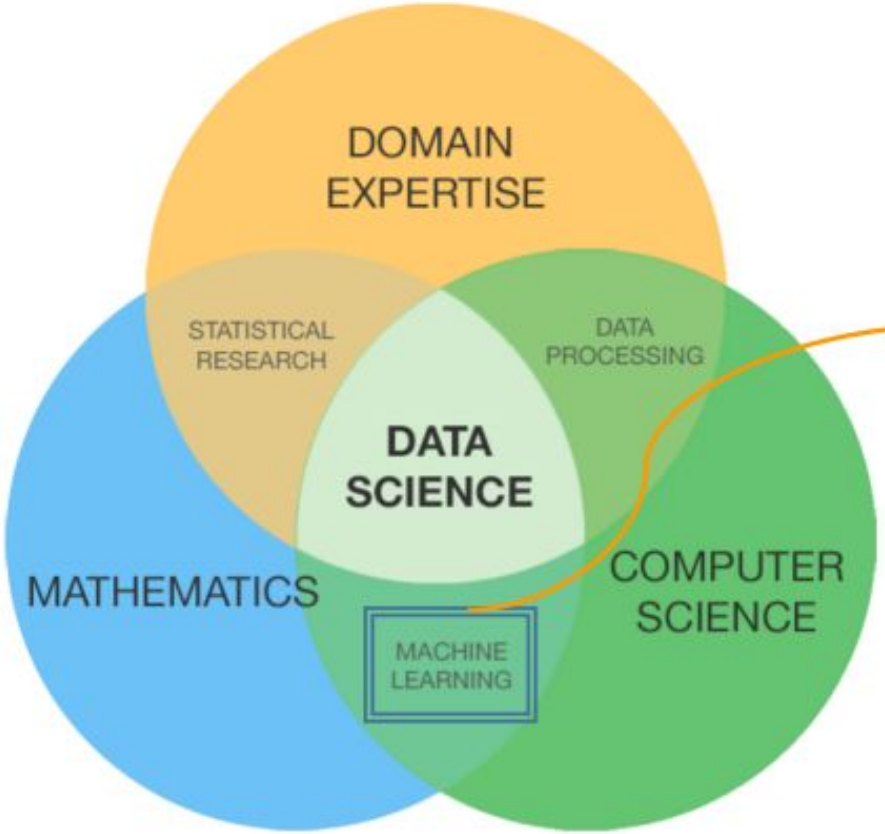
Presentación

Guillermo González Sánchez

Ciencia de datos, *devops* y arquitectura

- https://www.linkedin.com/in/guillermogsds
- https://github.com/Guillermogsjc

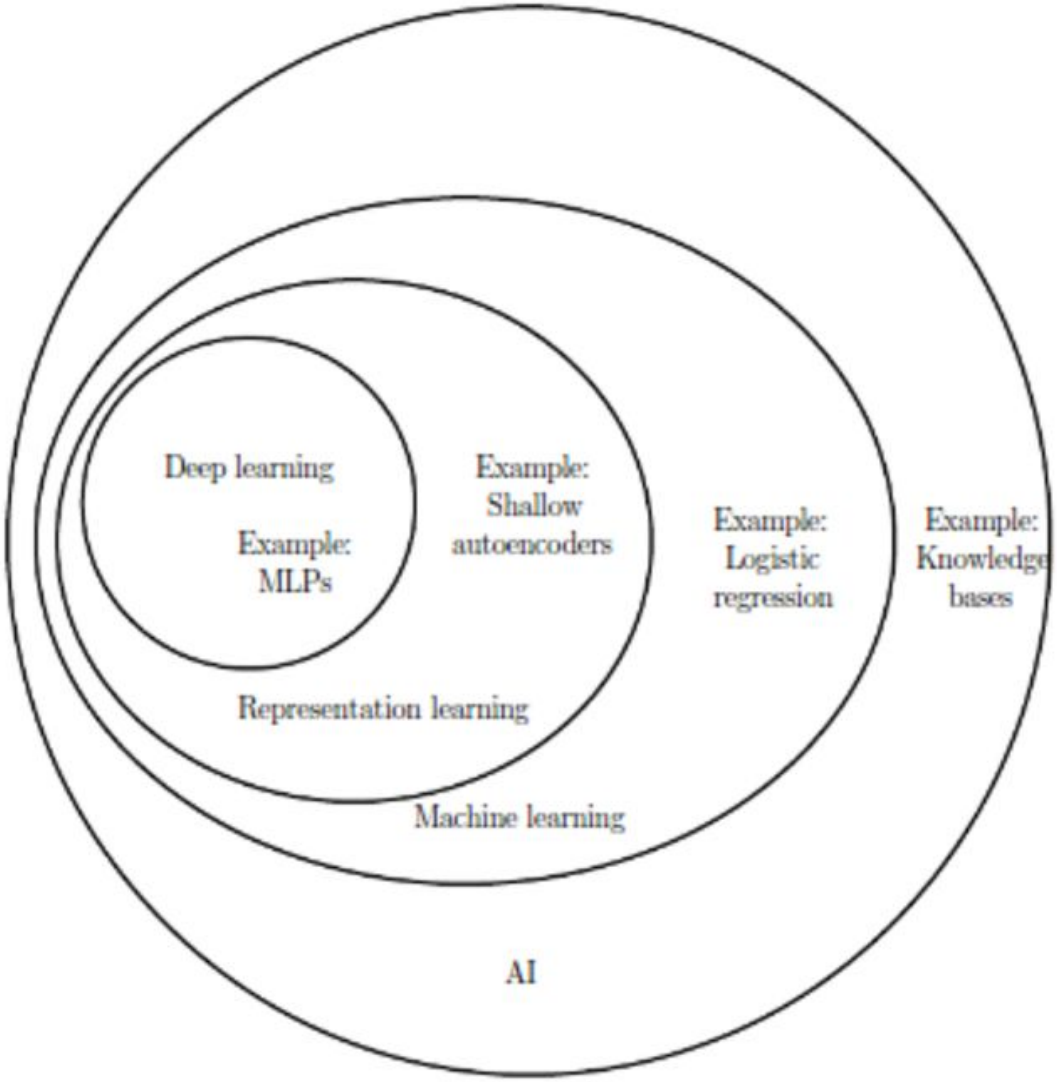
Licenciado en Matemáticas



- Matemáticas, estadística y algoritmia.
- Programación
- Conocimiento de dominio

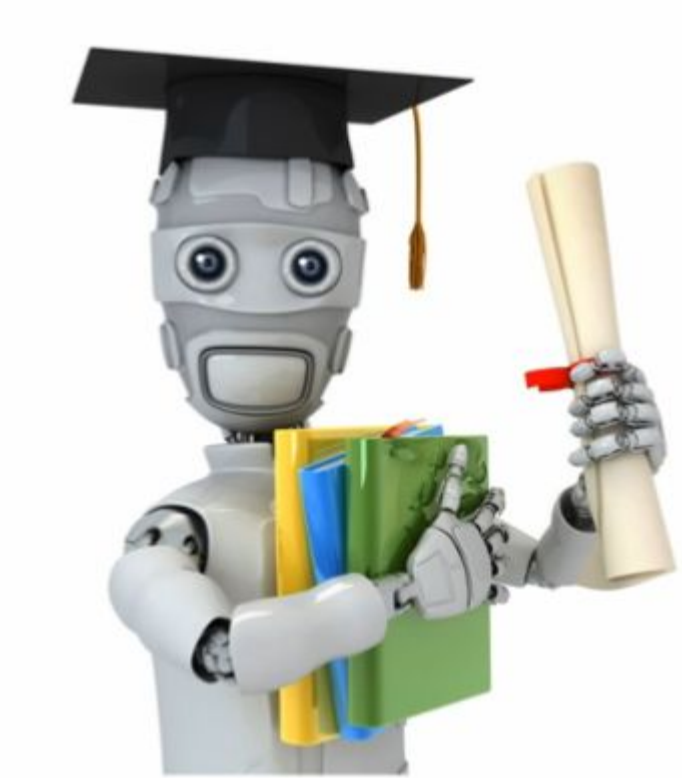
- **Aprendizaje automático o Machine Learning es una rama de la inteligencia artificial.**
- **Engloba un conjunto de técnicas que permite a las computadoras aprender patrones (o descubrirlos) basándose en la experiencia previa, esto es en los datos.**
- **Su asociación con big data es inmediata, si tenemos datos que hemos recolectado durante el ejercicio de nuestra actividad empresarial, podemos utilizar este conjunto de técnicas para construir sistemas de aprendizaje automático**
- **Hay un cambio de paradigma**
  - BI tradicional responde a qué ha ocurrido.
  - Big data (normalmente queremos referirnos a analítica avanzada). **Qué va a ocurrir**

Aprendizaje Automático como rama de la inteligencia artificial



Machine Learning (en adelante ML) consiste en un proceso de
inducción de conocimiento a través de un programa que es capaz
de generalizar comportamientos a partir de una información no
estructurada en forma de ejemplos.

En el proceso se genera un modelo que aprende de los datos y nos
sirve bien para clasificar o predecir, o bien para conocer alguna
característica de los datos que antes no conocíamos.

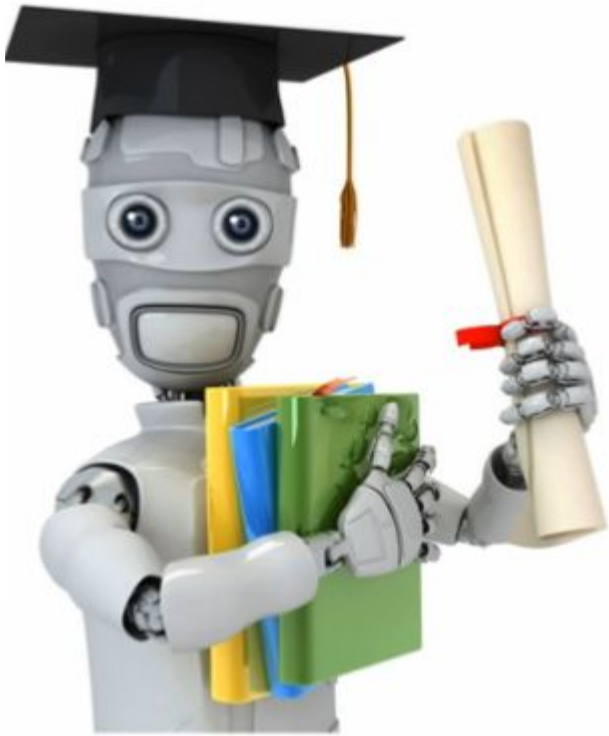


- **Arthur Samuel(1959):**

  Campo de estudio que permite a las computadoras aprender sin ser explícitamente programadas.

- **Tom Mitchell(1998):**

  Problema de ML bien planteado:
  Un programa se dice que ***aprende*** *de la experiencia E respecto a alguna tarea T y alguna medida de eficacia P si su eficacia en T, medida por P, mejora con la experiencia E.*

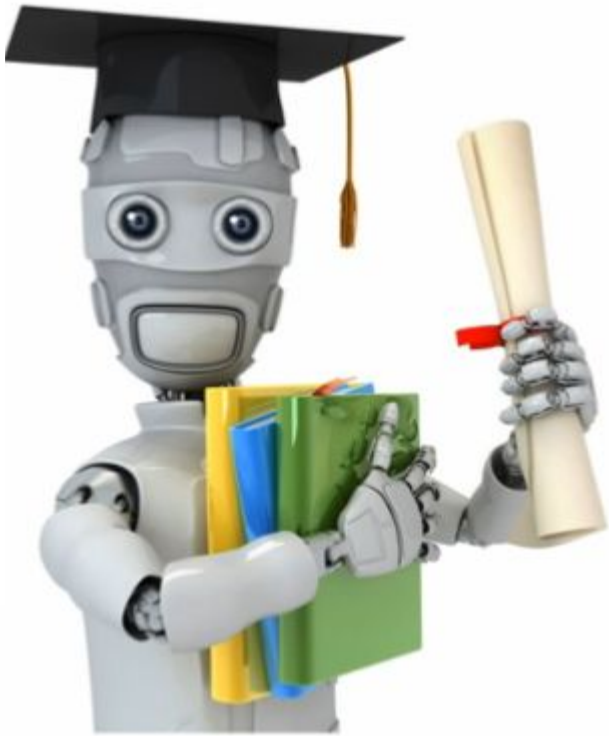


- **Arthur Samuel(1959):**

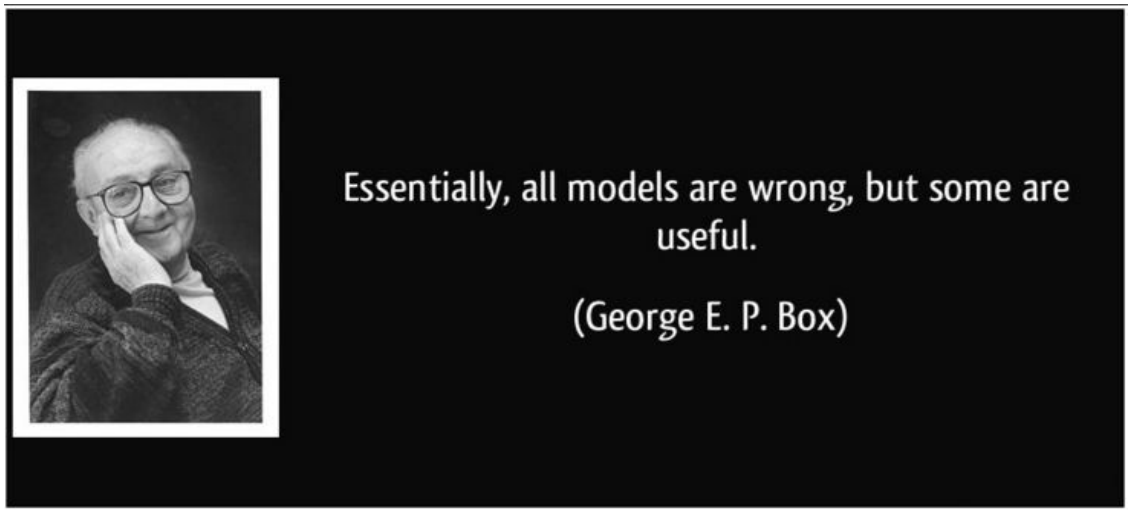
  Campo de estudio que permite a las computadoras aprender sin ser explícitamente programadas.

- **Tom Mitchell(1998):**

  Problema de ML bien planteado:
  Un programa se dice que ***aprende*** *de la experiencia E respecto a alguna tarea T y alguna medida de eficacia P si su eficacia en T, medida por P, mejora con la experiencia E.*

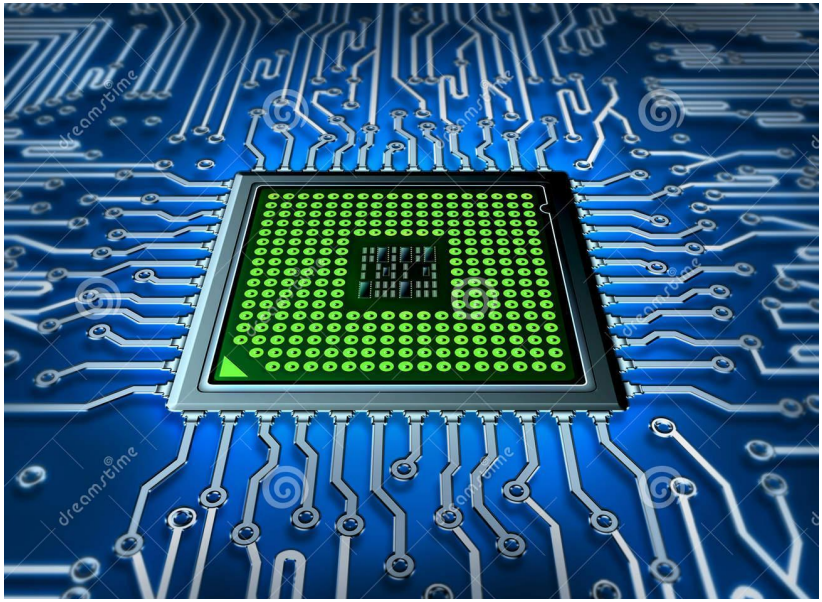


El objetivo del aprendizaje supervisado es hacer predicciones a futuro basadas en comportamientos o características que se han visto en los datos ya almacenados (el histórico de datos), es decir crear modelos predictivos.



El objetivo del científico de datos es construir modelos que mejor se aproximen a la realidad del negocio

¿Por qué el momento es ahora? ¿Data Mining de siempre?



Las empresas llevan tiempo almacenando datos

- Apriori
- MapReduce
- Association rules
- Frequent itemsets
- PCY
- Recommender systems
- PageRank
- TrustRank
- HITS
- SVM
- Decision Trees
- Perceptron
- Web Advertising
- DGIM
- Bandits
- BFR
- Regret
- LSH
- MinHash
- SVD
- Clustering
- Matrix factorization
- CUR
- Bloom filters
- Flajolet-Martin
- CURE
- Submodularity
- SGD
- Collaborative Filtering
- SimRank
- Random hyperplanes
- Trawling
- AND-OR constructions
- k-means

$$G = \left| 1 - \sum_{k=1}^{k=n-1} (X_{k+1} - X_k)(Y_{k+1} + Y_k) \right|$$

Pero ahora hay suficiente CAPACIDAD DE COMPUTACIÓN

Los algoritmos existen desde hace mucho

Ejemplos:

- Búsqueda web
- Etiquetado de fotos
- Identificación de usuario de tarjeta de crédito
- Sistemas de recomendación
- Filtros de spam en los correos

Funciones:

- Crear aplicaciones que no se podrían generar a mano
- Entender la gran cantidad de información y datos generados hoy día
- Programas que se personalizan solos en función de los datos de clientes
- Realizar predicciones y clasificaciones sobre nuevas entradas de datos

Conducción automática con redes neuronales

https://www.youtube.com/watch?v=ilP4aPDTBPE

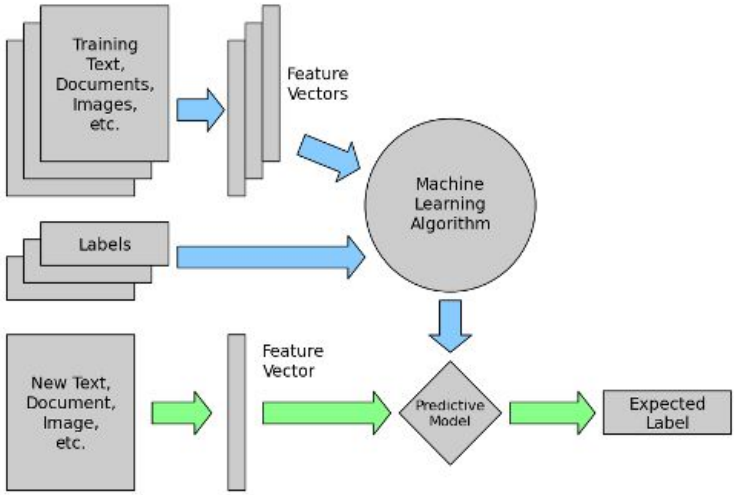
Dilemas morales

http://moralmachine.mit.edu/

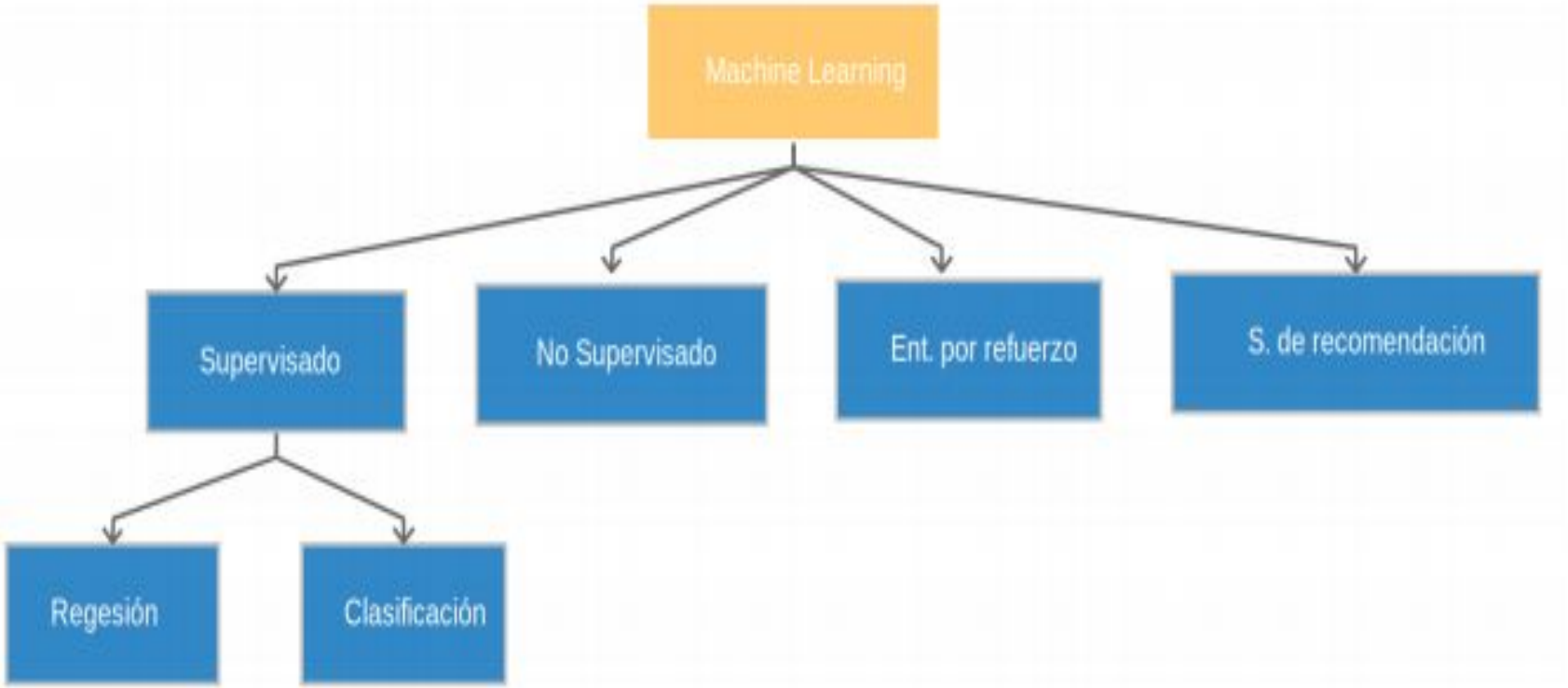
- Identificar el objetivo de negocio
- Identificar las fuentes de datos
- Construir el conjunto de datos

  a) Desarrollar el modelo
  b) Medir resultado con el objetivo de nego
  c) Identificar opciones de mejora
  d) Implementar las mejoras para mejorar e modelo

- Ponerlo en producción



Tipos de aprendizaje automático



Tipos de aprendizaje automático

- Aprendizaje supervisado: **Se basan en la existencia de un conjunto de datos de entrenamiento que contienen la respuesta correcta al problema. Por ejemplo si queremos predecir el precio de una casa el conjunto de datos de entrenamiento contiene el precio de las casas.**

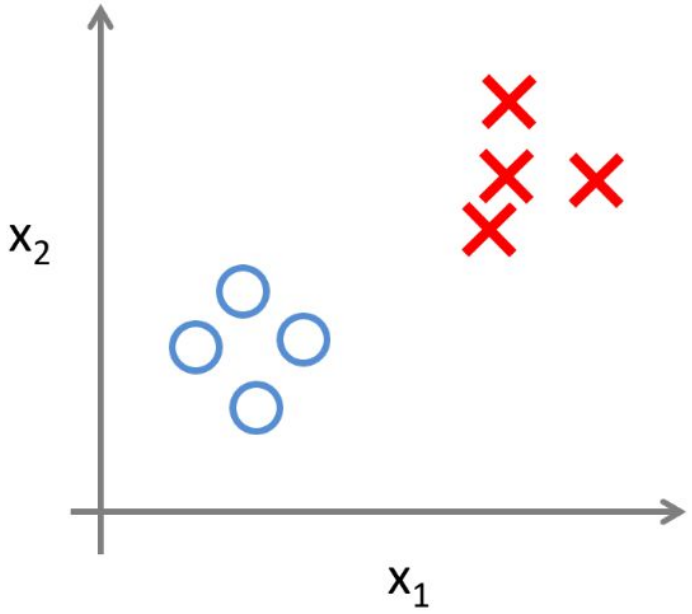
  EJEMPLOS ETIQUETADOS -> ENTRENAMIENTO ALGORITMO -> MODELO PREDICTIVO

- Aprendizaje no supervisado: **Se basan en la existencia de un conjunto de datos de entrenamiento que NO contienen la respuesta correcta al problema. Si misión u objetivo es buscar estructura en los datos.**

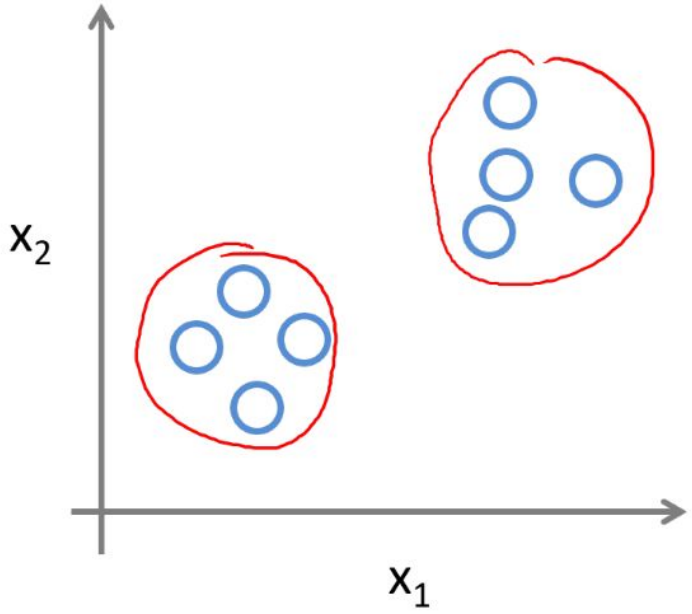
  EJEMPLOS -> ENTRENAMIENTO ALGORITMO -> ESTRUCTURA SOBRE LOS EJEMPLOS

- Aprendizaje por refuerzo: **Hay una serie de posibles acciones a tomar en una secuencia con distintas recompensas según el contexto y se quiere desarrollar un agente que optimice el proceso maximizando la suma de las recompensas obtenidas.**

Supervised Learning



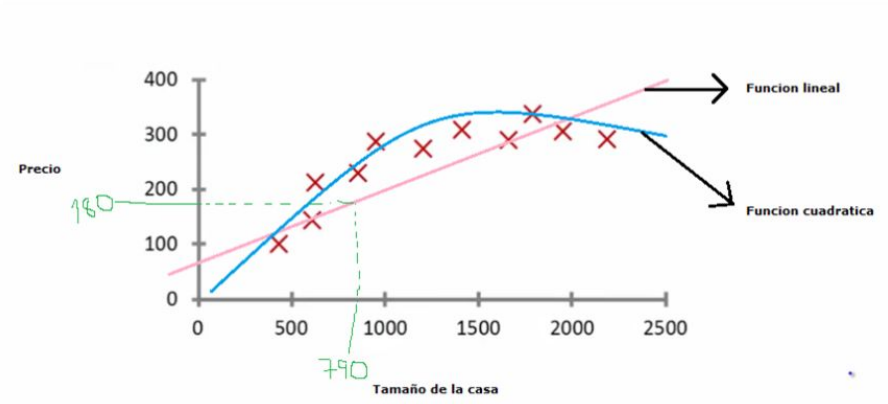
Unsupervised Learning



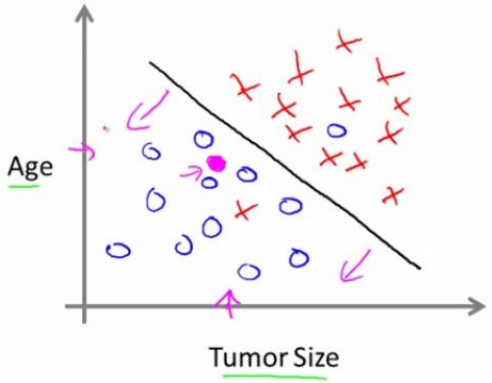
Tenemos dos grandes grupos de problemas para los que podemos usar aprendizaje supervisado:

- **Regresión: Cuando nuestro sistema quiere predecir una variable continua. Ejemplo el precio de una casa, las ventas de un producto.**

- **Clasificación: Cuando nuestro sistema quiere predecir la clase o el tipo al que pertenece una observación. Es decir predecimos una variable discreta.**

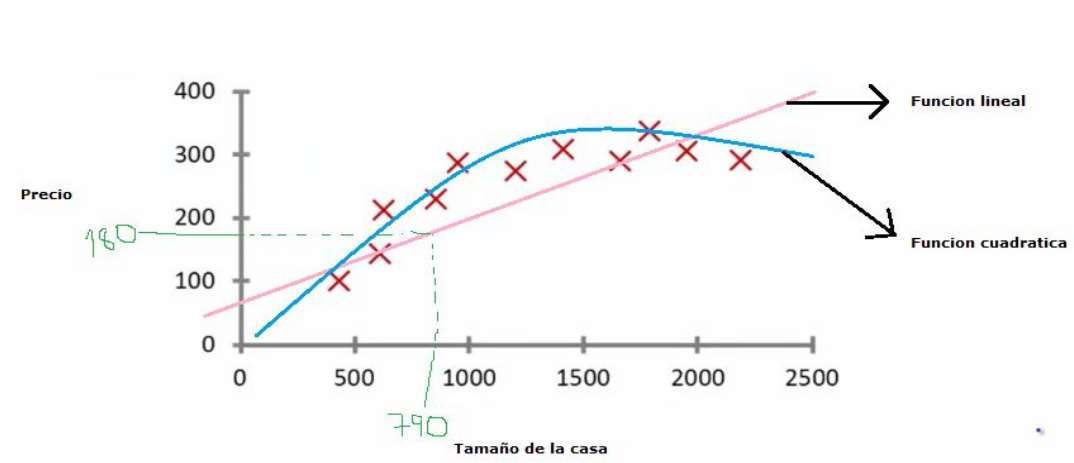


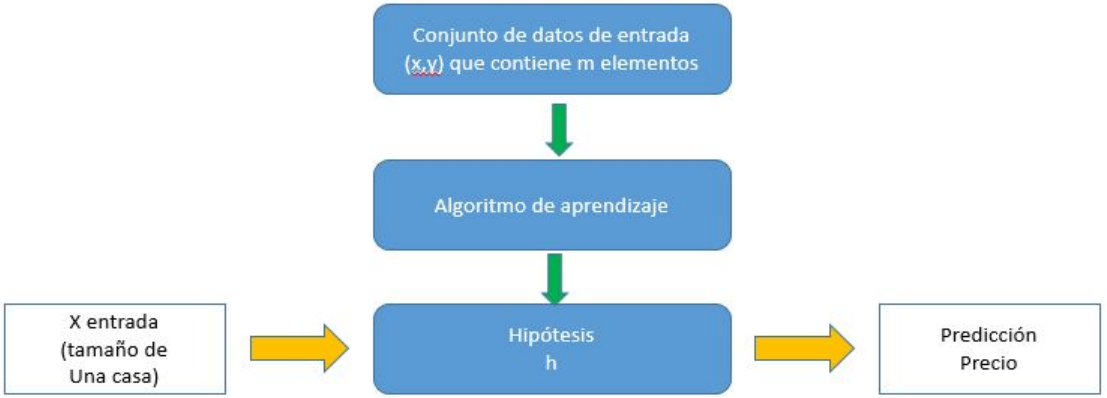
(a) Regresión

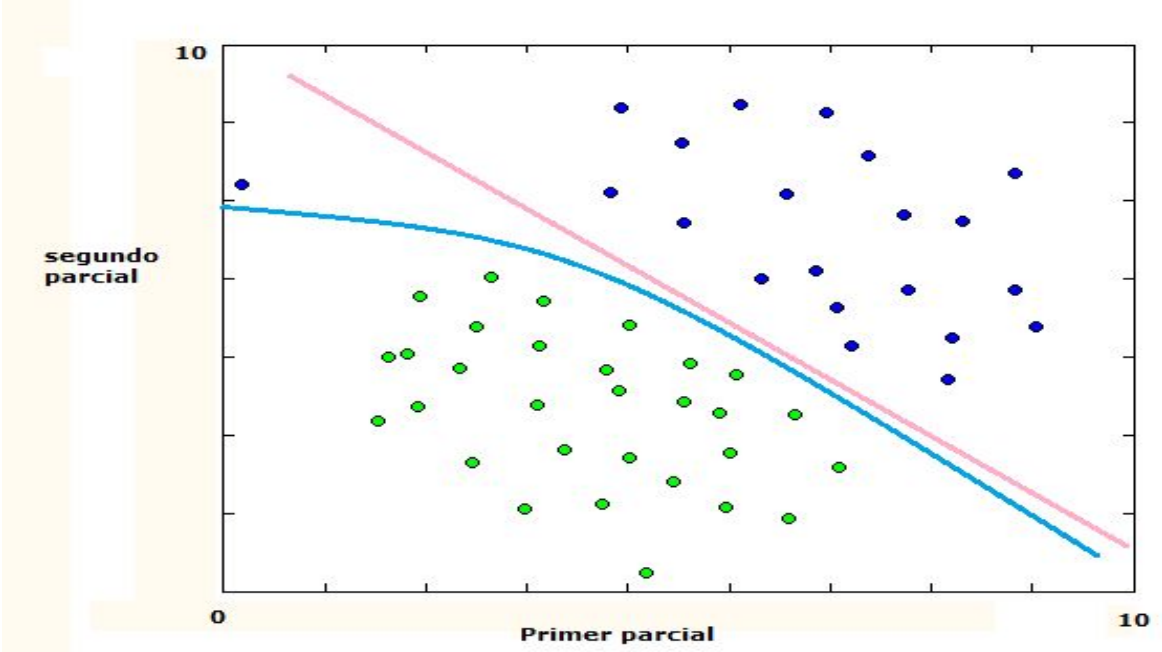


(b) Clasificación

A partir de unas características, como puede ser el tamaño de una casa, intentamos predecir el precio de la misma.

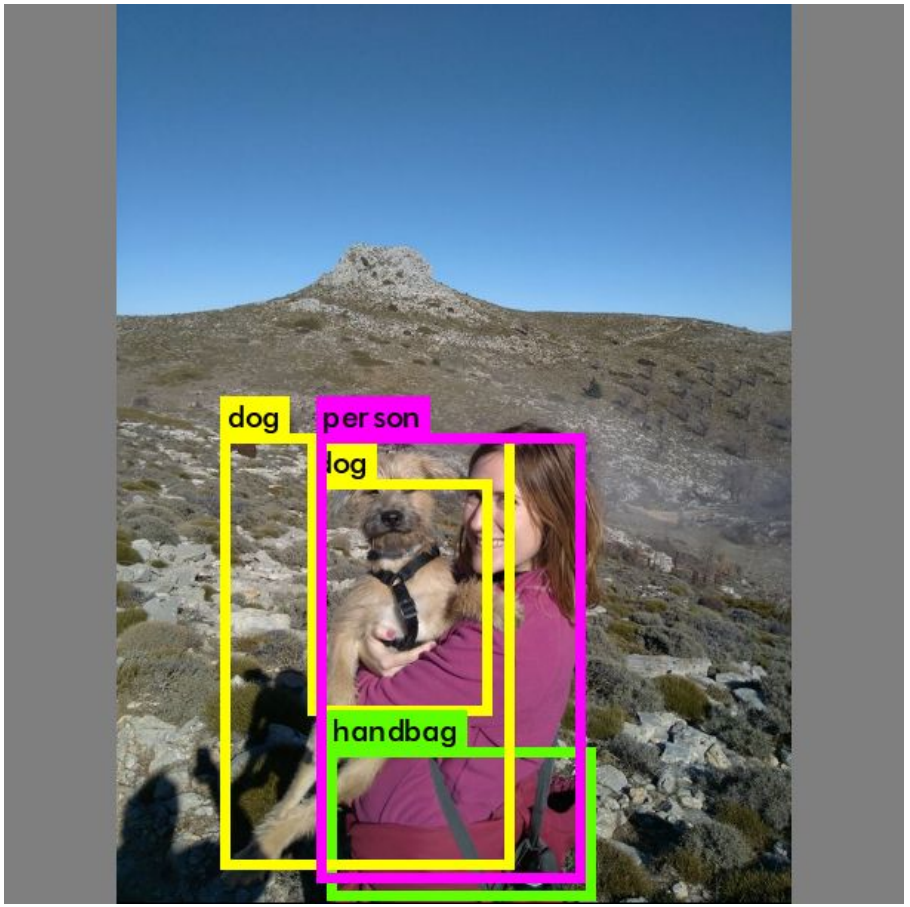






A partir de unas características, como puede ser la nota en el primer parcial y la nota en el segundo parcial queremos saber qué alumnos aprobarán la asignatura.

A partir de las características de una foto identificar el tipo de objeto que hay



Vemos el vídeo de etiquetado en tiempo real

https://pjreddie.com/darknet/yolo/

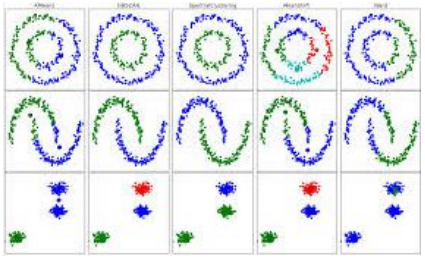
- **Modelo de detección fraude en reservas para agencia de viajes**
- **Modelo de cancelaciones para empresa hotelera**
- **Modelo de recomendación de productos complementarios para empresa hotelera**
- **Modelo de mantenimiento predictivo para industria de minería**
- **Modelo de asignación de plazas libres en zonas de estacionamiento regulado**
- **Modelo de verificación de firmas en cheques para entidad bancaria**
- **Modelo de detección de averías en cables de tensión con drones**
- **Modelo de segmentación y detección de anomalías en reparaciones de vehículos**
- **Producto de analítica conversacional completo (S2T, NLU, NLP)**

Afianzamos conceptos, los siguientes ejemplos son regresión o clasificación:

- **Si queremos predecir el número de pantalones del tipo X que vamos a vender en el próximo mes.**
- **Si queremos predecir el precio de venta de una casa.**
- **Si queremos predecir si una maquina va a fallar por un error.**
- **Si queremos predecir si un cliente va a dejarnos.**
- **Si queremos predecir la nota de un alumno en un examen. ¿Y si va a suspender ó no?**
- **Identificar si un correo es spam o no es spam.**
- **Si queremos saber el numero de apuestas que una empresa va a recibir para un evento**
- **Si queremos saber el tipo de apuestas (digamos que son dos) que vamos a recibir**

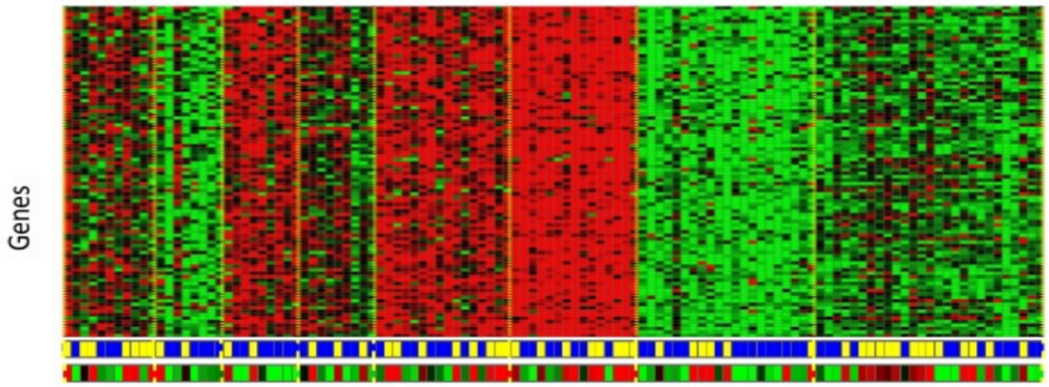
Se basan en la existencia de un conjunto de datos de entrenamiento que NO contienen la respuesta correcta al problema. Si misión u objetivo es buscar estructura en los datos.

- **Los más usados son los algoritmos de segmentación ó clustering, buscan una estructura para agrupar los datos en conjuntos en donde los integrantes de cada uno son similares entre sí en base a un criterio ó distancia**



- **PCA. Es un algoritmo de reducción de la dimensionalidad. Buscan una superficie en la que proyectar los datos minimizando la perdida de información.**

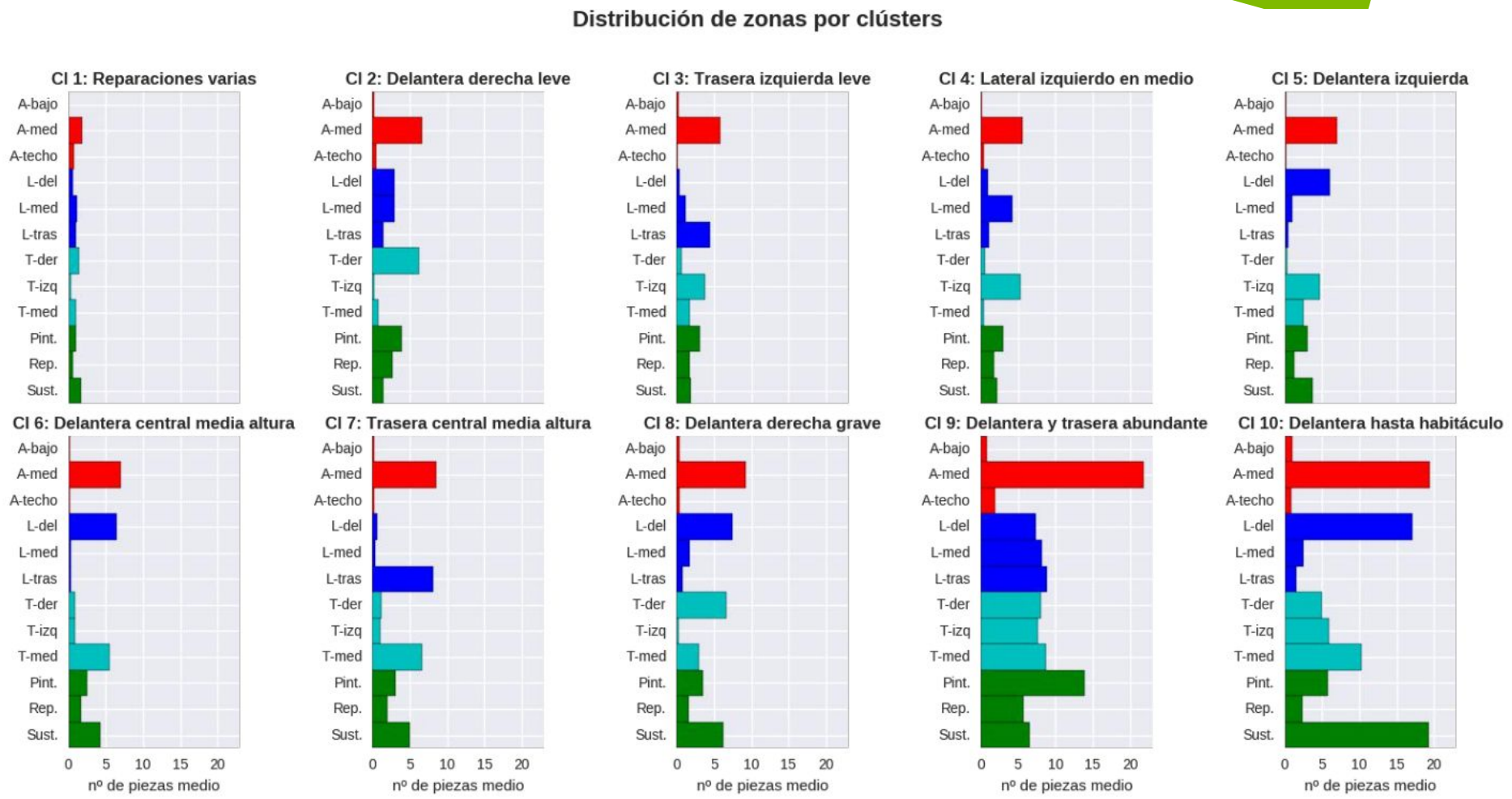
Forma parte del proceso de búsqueda de conocimiento ya que se hace sobre datos no etiquetados



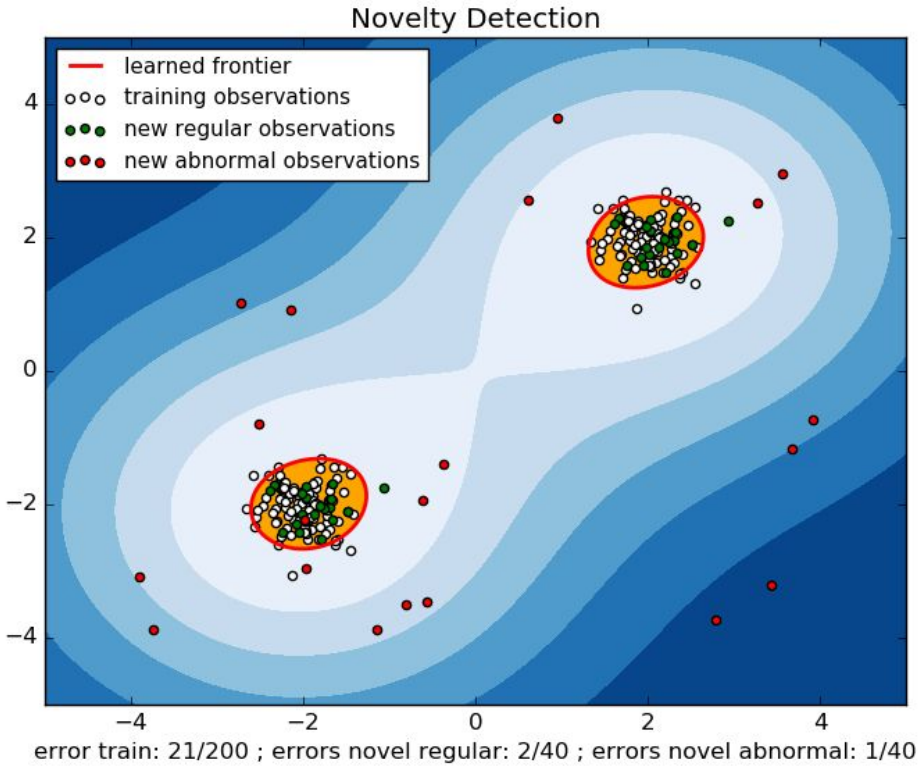
DNA microarray data to understand genomics

Colors show the degree to which different individuals do or do not have a specific gene.

Aprendizaje no supervisado: ejemplo SBD



Aprendizaje no supervisado en IOT: *novelty detection*

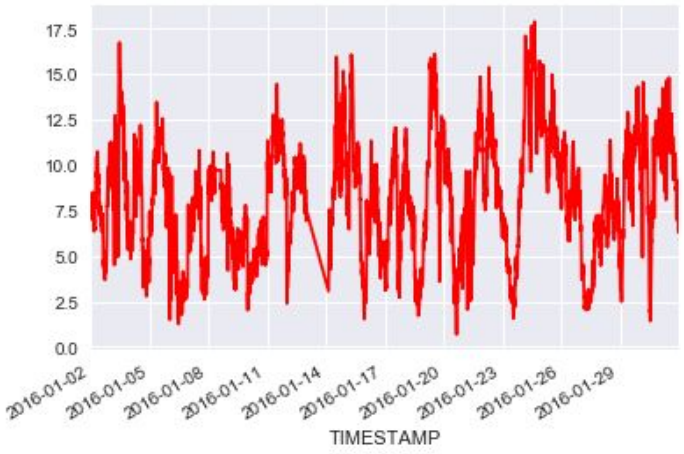


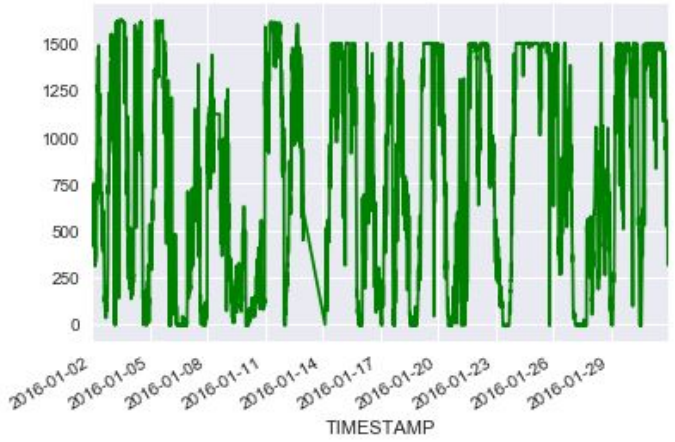
Consiste en identificar casos nuevos que salen de la rutina de funcionamiento de una máquina. Estos se identifican como potenciales anomalías para su estudio.
Cualquier máquina con sensores produce datos contínuamente y es aplicable este tipo de análisis.

CASO DGT: Reconocimiento de características de accidentes

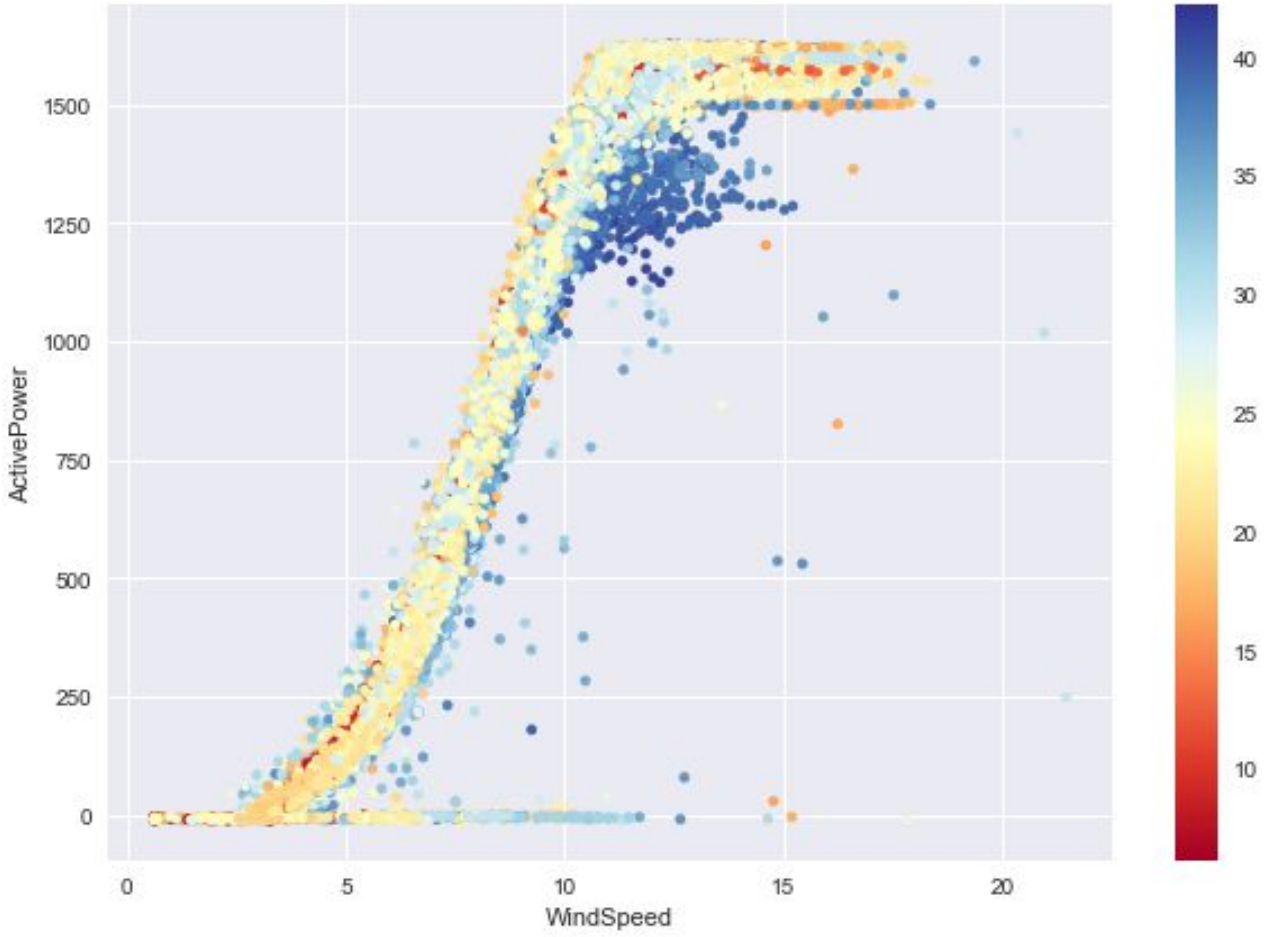
Análisis de molinos en planta eólica:

TIMESTAMP	WindSpeed_01	ActivePower_01	Shaft_Bearing_01	WindSpeed_02	ActivePower_02	Shaft_Bearing_02
2016-01-14 03:10:00	3.059544	-4.814381	22.380674	0.099995	-4.156104	22.224743
2016-01-14 03:20:00	4.117228	73.316517	22.386155	0.099996	-4.181894	22.254850
2016-01-14 03:30:00	5.361675	142.541678	22.424966	0.100005	-4.356417	22.146528
2016-01-14 03:40:00	5.955890	204.728167	22.410335	3.005707	97.889972	22.201758
2016-01-14 03:50:00	6.396137	278.722572	22.457345	4.554544	213.704642	22.162527





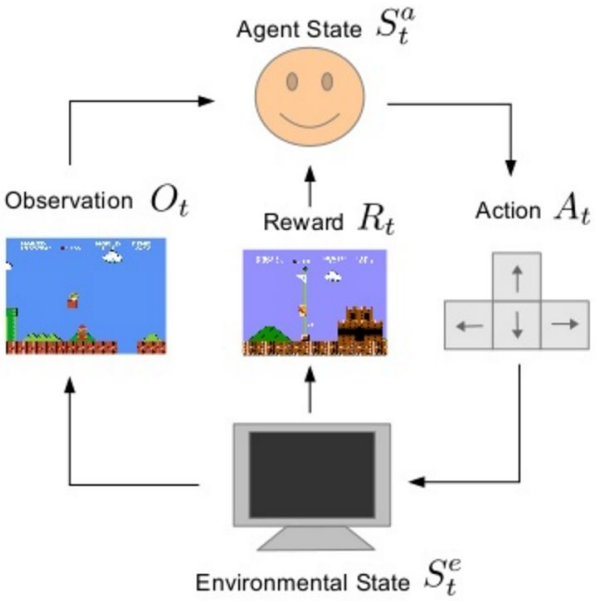
Análisis de molinos en planta eólica:



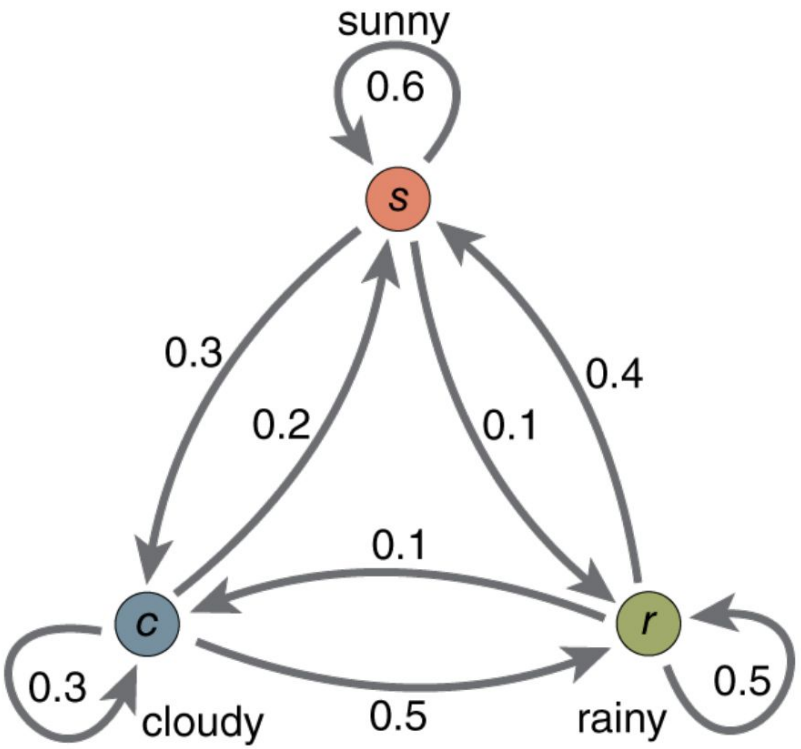
Hay un entorno en el que se tienen una serie de estados en los que se pueden realizar una cantidad de acciones finitas. Tras cada acción se obtiene una recompensa.

Hay que seleccionar para cada estado cuál es la acción que da una mayor recompensa.

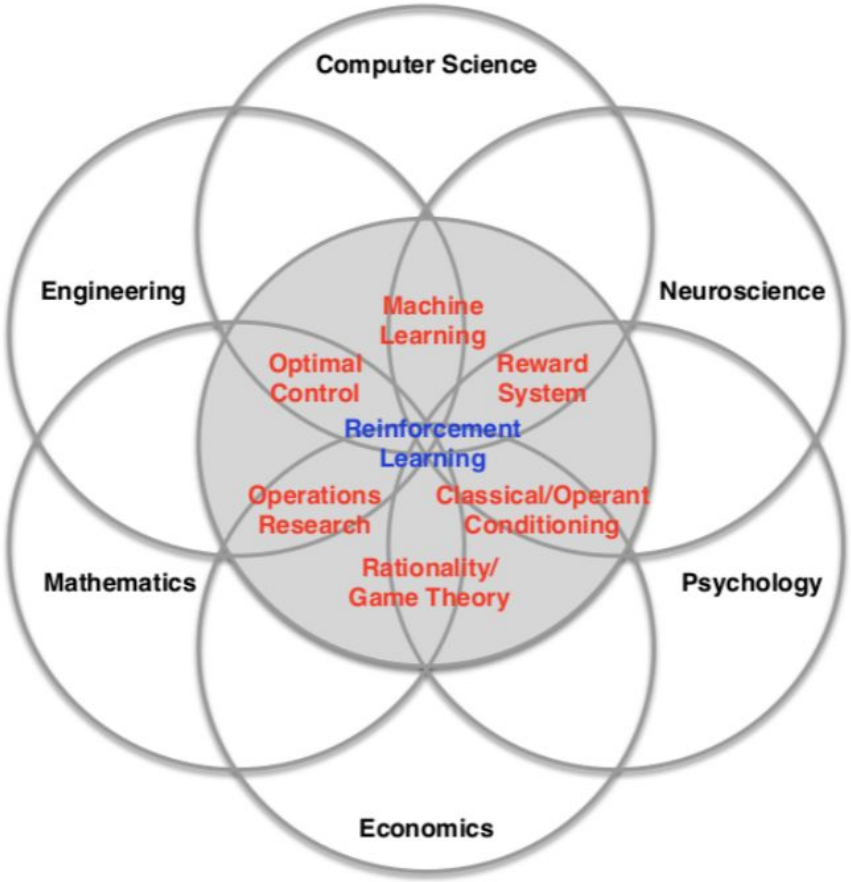
https://www.youtube.com/watch?v=YOW8m2YGtRg

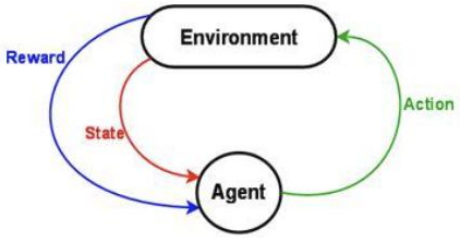


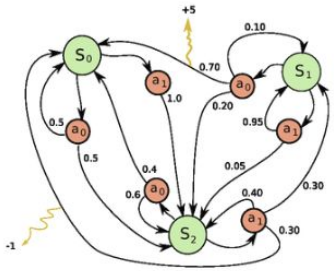
- Rules of the game are unknown.
- No supervisor, only a reward signal.
- Feedback is delayed.
- Agent's actions affect the subsequent data it receives.



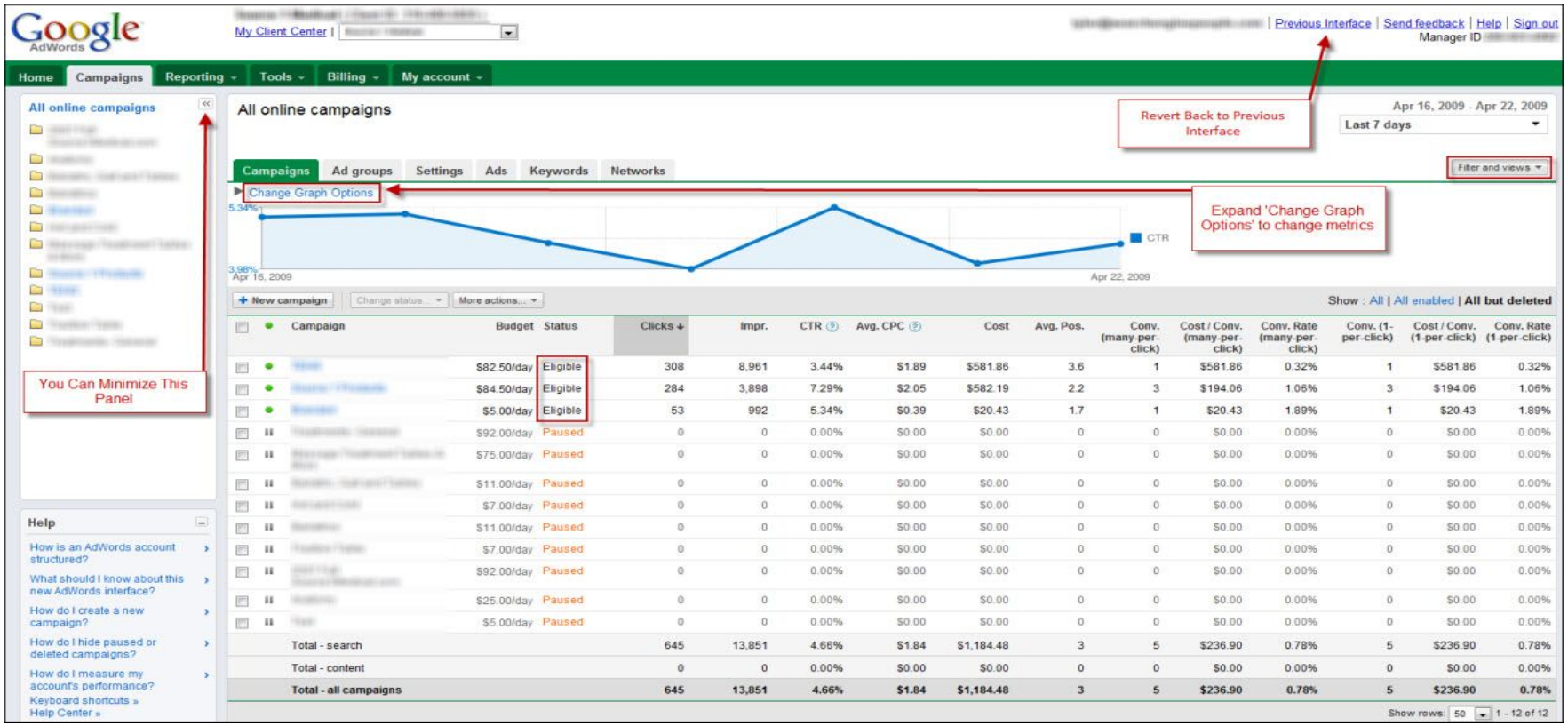
Aprendizaje por refuerzo







Aprendizaje por refuerzo: ejemplo SBD

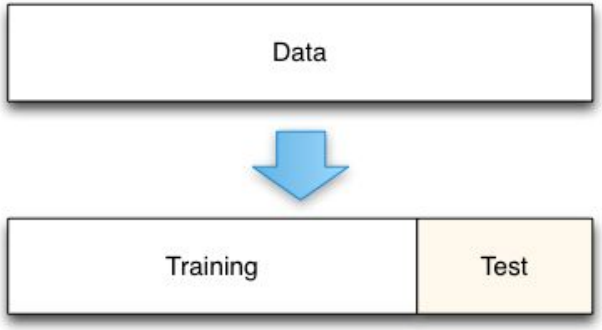


Campaign	Budget	Status	Clicks	Impr.	CTR	Avg. CPC	Cost	Avg. Pos.	Conv. (many-per-click)	Cost / Conv. (many-per-click)	Conv. Rate (many-per-click)	Conv. (1-per-click)	Cost / Conv. (1-per-click)	Conv. Rate (1-per-click)
	$82.50/day	Eligible	308	8,961	3.44%	$1.89	$581.86	3.6	1	$581.86	0.32%	1	$581.86	0.32%
	$84.50/day	Eligible	284	3,898	7.29%	$2.05	$582.19	2.2	3	$194.06	1.06%	3	$194.06	1.06%
	$5.00/day	Eligible	53	992	5.34%	$0.39	$20.43	1.7	1	$20.43	1.89%	1	$20.43	1.89%
	$92.00/day	Paused	0	0	0.00%	$0.00	$0.00	0	0	$0.00	0.00%	0	$0.00	0.00%
	$75.00/day	Paused	0	0	0.00%	$0.00	$0.00	0	0	$0.00	0.00%	0	$0.00	0.00%
	$11.00/day	Paused	0	0	0.00%	$0.00	$0.00	0	0	$0.00	0.00%	0	$0.00	0.00%
	$7.00/day	Paused	0	0	0.00%	$0.00	$0.00	0	0	$0.00	0.00%	0	$0.00	0.00%
	$11.00/day	Paused	0	0	0.00%	$0.00	$0.00	0	0	$0.00	0.00%	0	$0.00	0.00%
	$7.00/day	Paused	0	0	0.00%	$0.00	$0.00	0	0	$0.00	0.00%	0	$0.00	0.00%
	$92.00/day	Paused	0	0	0.00%	$0.00	$0.00	0	0	$0.00	0.00%	0	$0.00	0.00%
	$25.00/day	Paused	0	0	0.00%	$0.00	$0.00	0	0	$0.00	0.00%	0	$0.00	0.00%
	$5.00/day	Paused	0	0	0.00%	$0.00	$0.00	0	0	$0.00	0.00%	0	$0.00	0.00%
Total - search			645	13,851	4.66%	$1.84	$1,184.48	3	5	$236.90	0.78%	5	$236.90	0.78%
Total - content			0	0	0.00%	$0.00	$0.00	0	0	$0.00	0.00%	0	$0.00	0.00%
Total - all campaigns			**645**	**13,851**	**4.66%**	**$1.84**	**$1,184.48**	**3**	**5**	**$236.90**	**0.78%**	**5**	**$236.90**	**0.78%**

Show rows: 50 | 1 - 12 of 12

Cuando queremos usar ML en una tarea determinada, asociamos un *dataset* la misma.
Los modelos por lo general deben entrenarse para aprender de los datos.
De este modo, se divide el dataset en dos:

- **Entrenamiento (*train)*:** conjunto de datos sobre los que el modelo aprenderá. Se suele tomar entre el 60% - 80%.
- **Prueba (*test*):** conjunto de datos sobre los que se prueba la eficacia del modelo



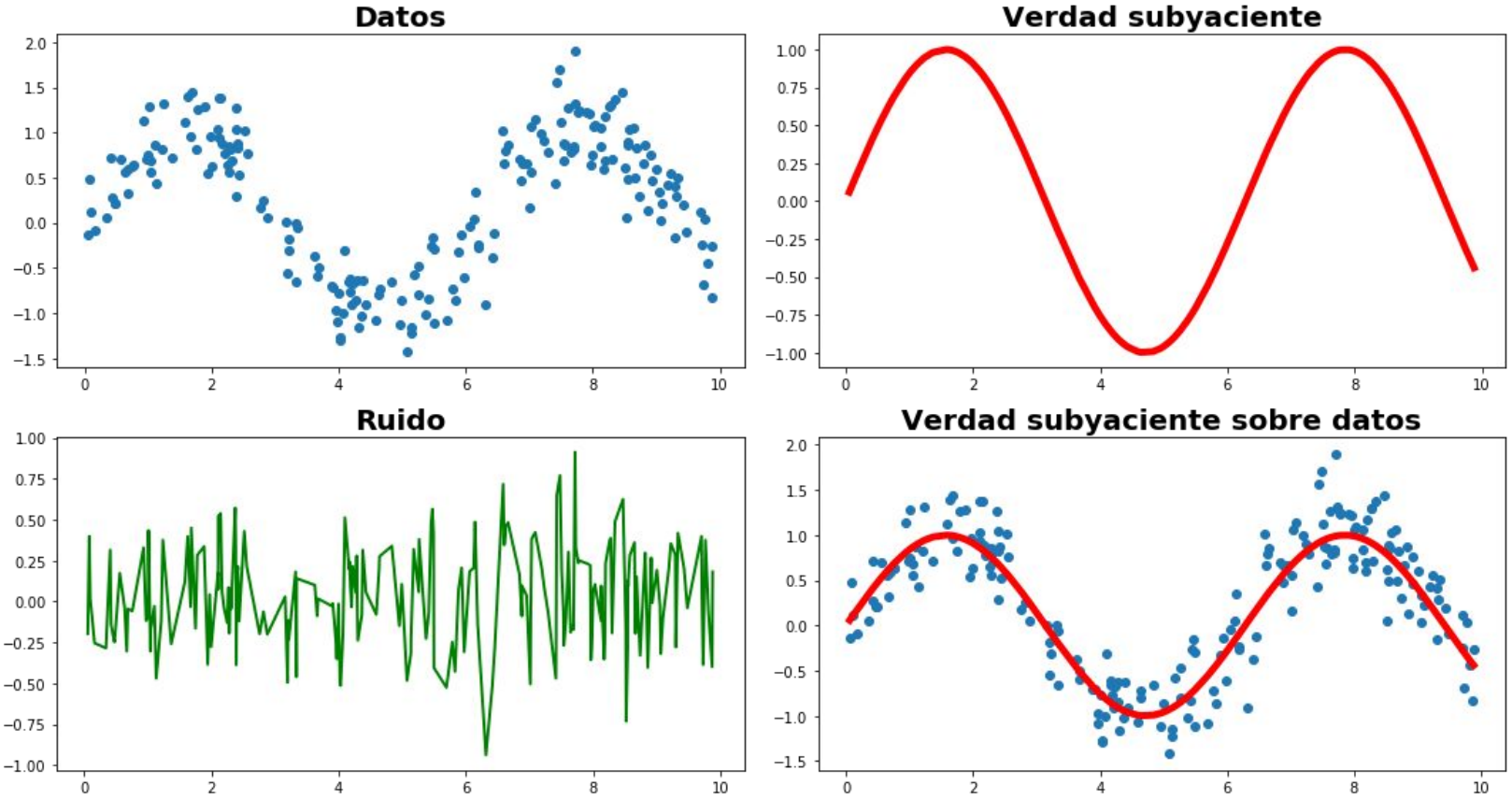
Además, algunos modelos requieren que en el conjunto de entrenamiento se tome una fracción para **validación**, este se usa para monitorizar el entrenamiento y saber cuándo se debe parar.

La evaluación del rendimiento de un modelo se realiza a través de las métricas tomadas del mismo en el conjunto de ***test*** .

Estos valores aproximan al rendimiento que se espera en la realidad cuando se aplique el modelo sobre datos nuevos

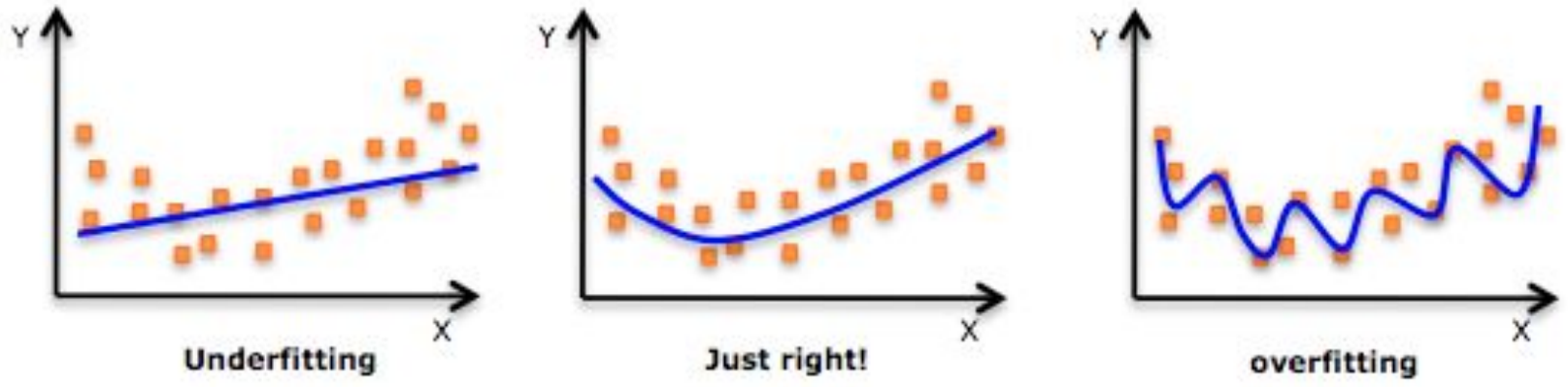
Para cada tipo de ML supervisado daremos una serie de métricas convenientes y cómo analizar los resultados.

El objetivo de un modelo de ML es captar los patrones que están contenidos en los datos, separando estos del **ruido**.
Se define **ruido** como las variaciones en los valores de los datos que no corresponden a ninguna regla ó ley que no sea la aleatoriedad.
Por tanto no estamos interesados en captar el ruido en los datos ya que no aporta información útil.



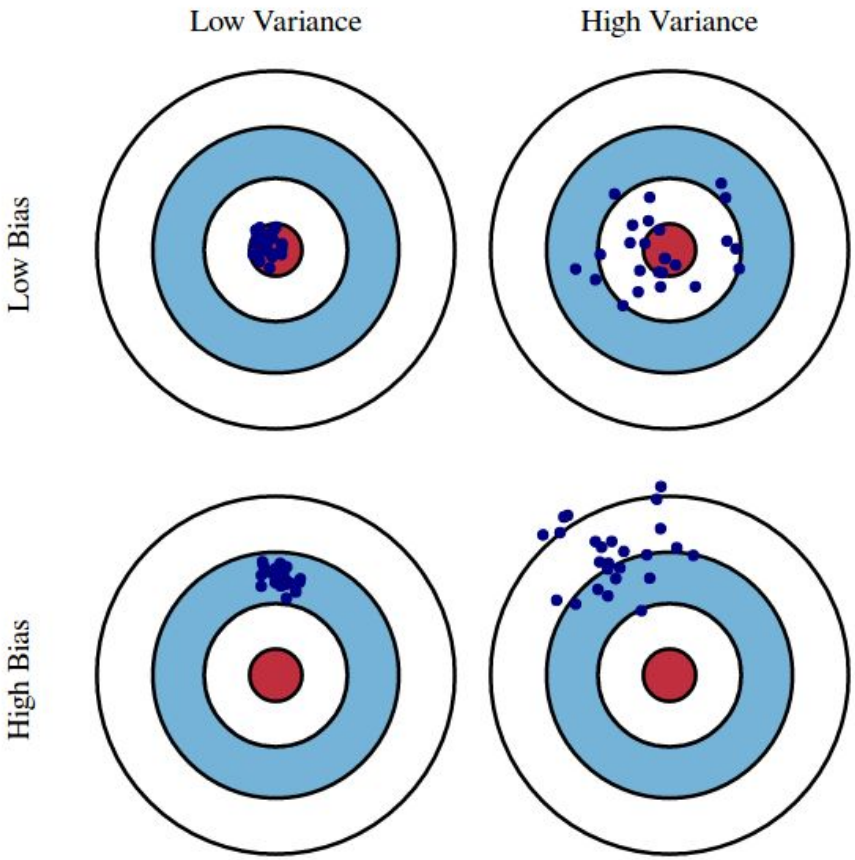
Cuando entrenamos un modelo sobre un conjunto de datos, este tiene una serie de hiperparámetros configurables que establecen su **complejidad**.

- Un modelo muy complejo aprende mucho de los datos de entrenamiento con el riesgo de ajustarse demasiado a ellos y no captar la verdad subyaciente.
  Estaremos en un escenario de ***overfitting.***
- Un modelo muy simple aprende poco de los datos de entrenamiento con el riesgo de no ser capaz de captar la verdad subyaciente.
  Estaremos en un escenario de ***underfitting.***



Equilibrio entre sesgo y varianza

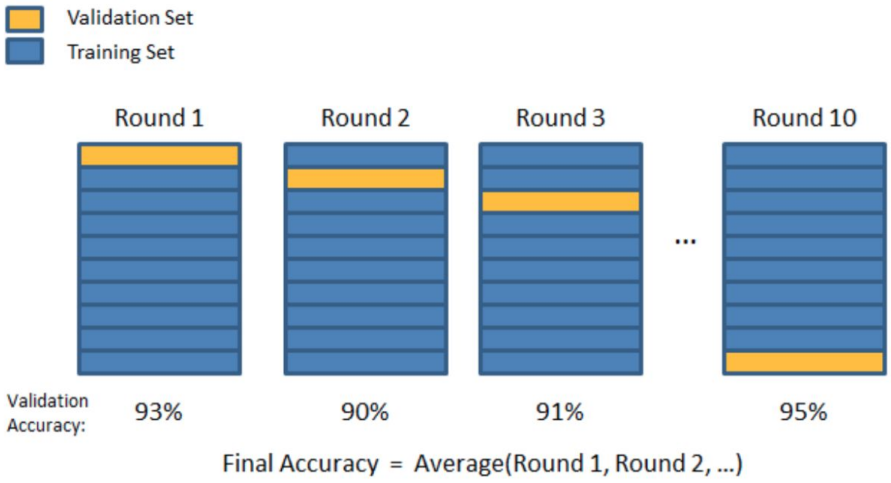
Cuando generamos un modelo estadístico sobre datos ó un modelo de ML, siempre tendremos que establecer un equilibrio entre sesgo y varianza según la capacidad de explicación del modelo ó su error de generalización.



Los modelos dependen de una serie de **hiperparámetros** posibles que indican valores de aplicación de los mismos a la hora de realizar el aprendizaje.

Para seleccionar estos hiperparámetros es conveniente probar suficientes combinaciones de los mismos y seleccionar la que mejor rendimiento aporta respecto a la métrica en la que valoraremos el rendimiento del modelo.

Un modo de hacerlo usando todos los datos disponibles es la validación cruzada en *k-folds*. Este consiste en dividir el *dataset* en k trozos y entrenar para cada parámetro k modelos en todos menos uno de los trozos, calculando el rendimiento en el trozo en donde no se realizó entrenamiento para cada modelo, y haciendo una media de los resultados de los mismos.



Los algoritmos de ML se entrenan minimizando una **función de error** que representa el valor de ajuste del modelo sobre los datos y se toma acorde al tipo de problema enfrentado.

Para minimizar esta función, por lo general se usan **algoritmos de optimización**.

Estos algoritmos suelen tener las cualidades:

- **Iterativo:** genera una secuencia de valores que convergen a un mínimo (al menos a un mínimo local)
- **Voraz:** en cada paso de la secuencia el algoritmo se mueve en la dirección en la que localmente se "desciende" más.

El ejemplo más extendido de este tipo de algoritmos es el de **gradiente decendiente**, presente en la gran mayoría de los entrenamientos de los modelos de ML.

Algoritmos de optimización: gradiente descendiente