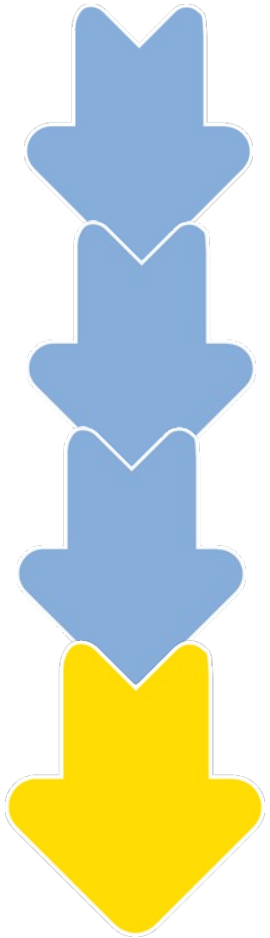
Aprendizaje automático

Clase 1

Iris Sattolo, Maximiliano Beckel

UNAHUR

00-Modelado y descubrimiento del conocimiento en Python



01 **Python**

02 **Procesamiento y exploración de datos**

03 **Aprendizaje automático**

04 **Modelado y aprendizaje no supervisado**

UNAHUR

00-Modelado y descubrimiento del conocimiento en Python

01 **Python**

02 **Procesamiento y exploración de datos**

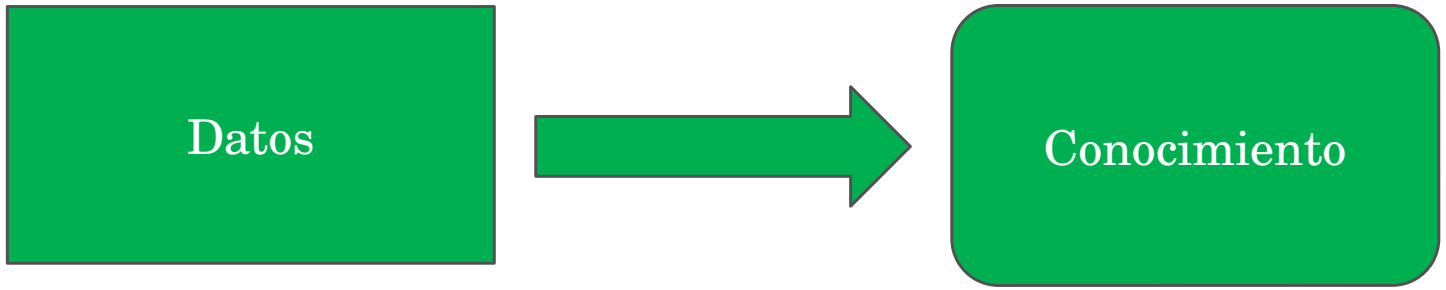
03 **Aprendizaje automático**

04 **Modelado y aprendizaje no supervisado**

UNAHUR

01-Introducción

El principal objetivo siempre es transformar datos en **conocimiento**. Y lo importante de este paso es que el conocimiento nos va a permitir tomar mejores decisiones!



01-Introducción

Aprendizaje Automático

Samuel (1959)
Campo de estudio que le da a las computadoras la habilidad de aprender sin ser programadas de manera explícita.

Mitchell (1998)
Un programa de computadora se dice que aprende de una **experiencia** E con respecto a una clase de **tareas** T y una medida de **performance** P, si su performance en las tareas T, medidas por P, mejoran con la experiencia E.

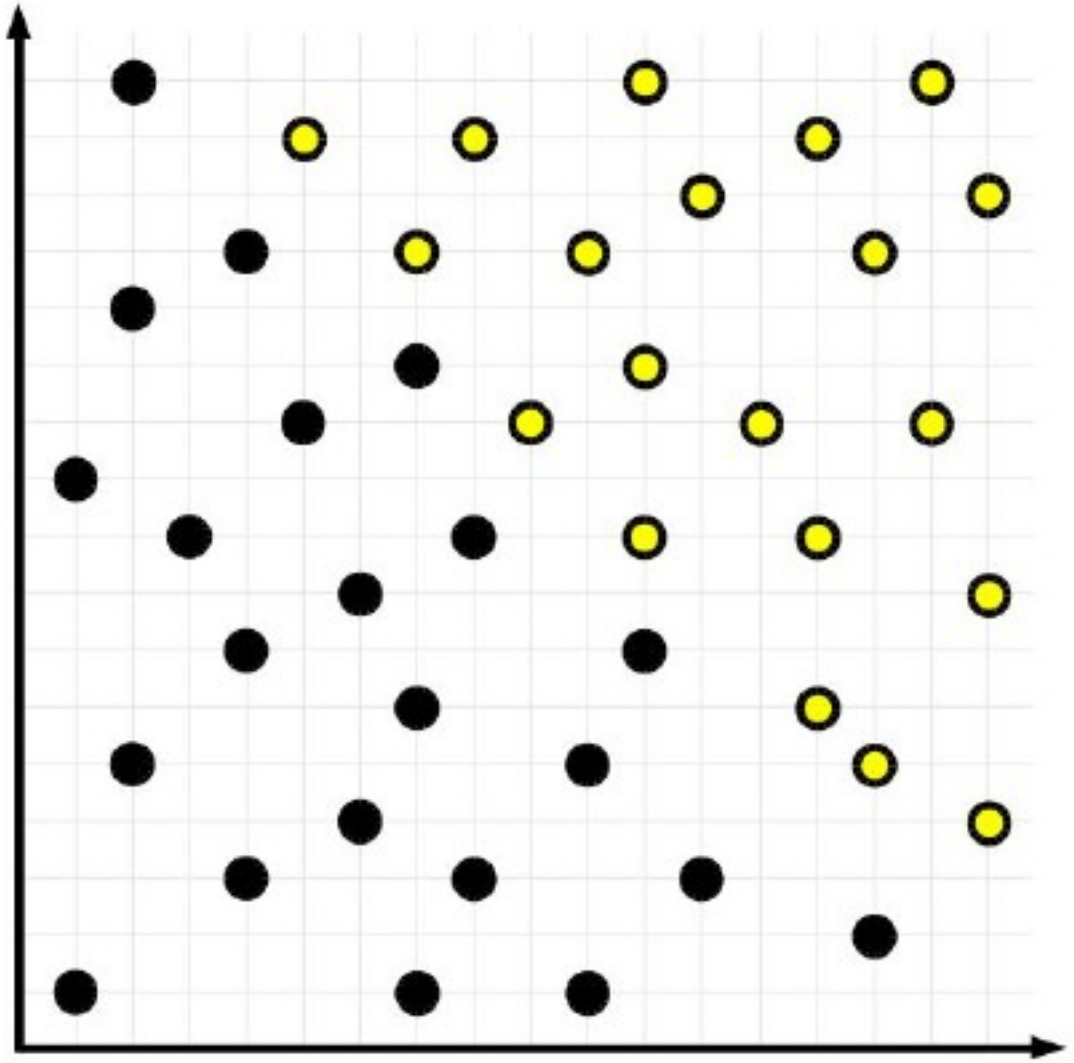
02-Aprendizaje Automático

Aprendizaje Automático

Un programa aprende una tarea si su performance mejora con la experiencia, y el aprendizaje es automático porque no le enseñamos a realizar esa tarea de manera explícita.

Tenemos que definir…

Tarea: objetivo
Experiencia: datos, ejemplos
Medida de performance: distancia formal a mi objetivo

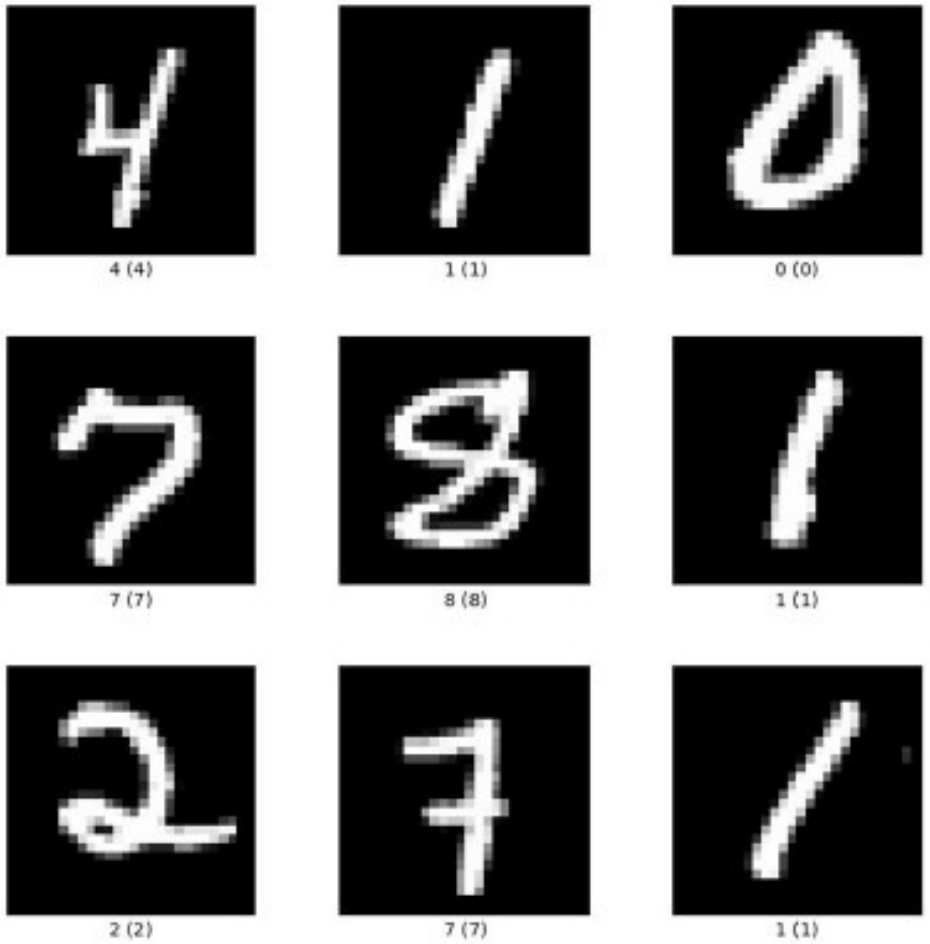


Aprendizaje Automático
Ejemplos

- **Tarea T**: predecir el color de un punto.
- **Experiencia de entrenamiento E**: Base de datos de puntos con sus respectivos colores.
- **Medida de desempeño P**: ¿qué porcentaje de puntos pude predecir correctamente su color?

UNAHUR

Aprendizaje Automático
Ejemplos



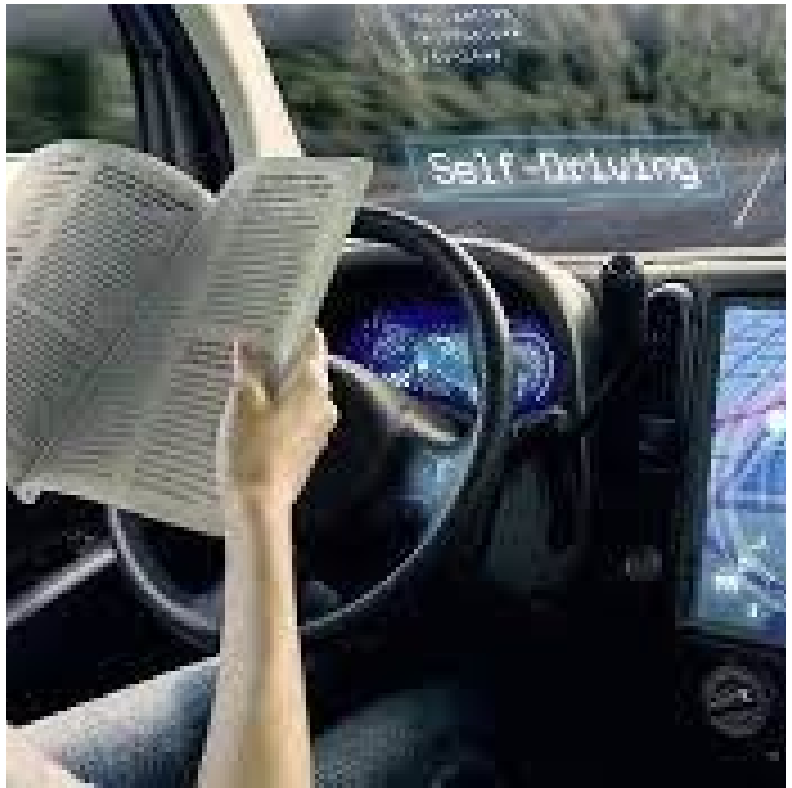
- **Tarea T**: digitalizar números escritos a mano.
- **Experiencia de entrenamiento E**: Imágenes de números a mano con sus etiquetas.
- **Medida de desempeño P**: % de digitalizaciones correctas.

UNAHUR

Aprendizaje Automático
Ejemplos

- **Tarea T**: Predecir el valor de una acción.
- **Experiencia de entrenamiento E**: Valor de la acción en el pasado.
- **Medida de desempeño P**: % predicciones correctas

UNAHUR



Aprendizaje Automático Ejemplo

- conducción autónoma

- **Tarea T**: conducción en carreteras públicas de cuatro carriles utilizando sensores de visión.
- **Experiencia de entrenamiento E**: una secuencia de imágenes y comandos de dirección registrados mientras se observa a un conductor humano.
- **Medida de desempeño P**: distancia promedio recorrida antes de un error (según lo juzgado por un supervisor humano)

UNAHUR

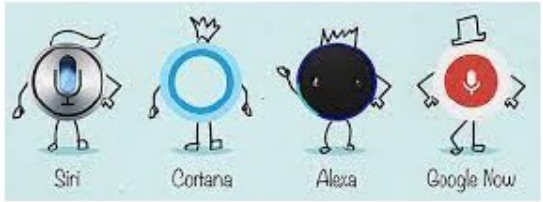
02-Aprendizaje automático

Sistemas de recomendación

Traductores

Predicción de tiempo de viaje, camino óptimo

Reconocimiento del habla, asistentes virtuales



Publicidad online, chatbots, text2image, detectar spam, reconocimiento de rostros/patentes, diagnósticos clínicos, vehículos autónomos, jugar al Go, …

02-Aprendizaje automático

MODELO: representación de una parte de la realidad que nos interesa entender con un determinado objetivo.

Modelar la realidad tiene sus limitaciones…

Sesgo Inductivo: conjunto de suposiciones que uno asume a la hora de construir un modelo a partir de mis datos.

- ➔ Tipo de modelo elegido
- ➔ Sesgo en nuestros datos

02-Aprendizaje automático

Tipos de Modelos según el tipo de Experiencia:

Aprendizaje supervisado:

Los datos están anotados con la respuesta correcta que quiero predecir.

- ➔ Clasificación: lo que quiero predecir es un clase (variable categórica)
- ➔ Regresión: lo que quiero predecir es un valor numérico.

Aprendizaje no supervisado:

Los datos de entrenamiento no están anotados.

- ➔ Encontrar patrones en nuestros datos.
- ➔ Agrupar nuestros datos en grupos homogéneos (clustering).

Aprendizaje por refuerzos:

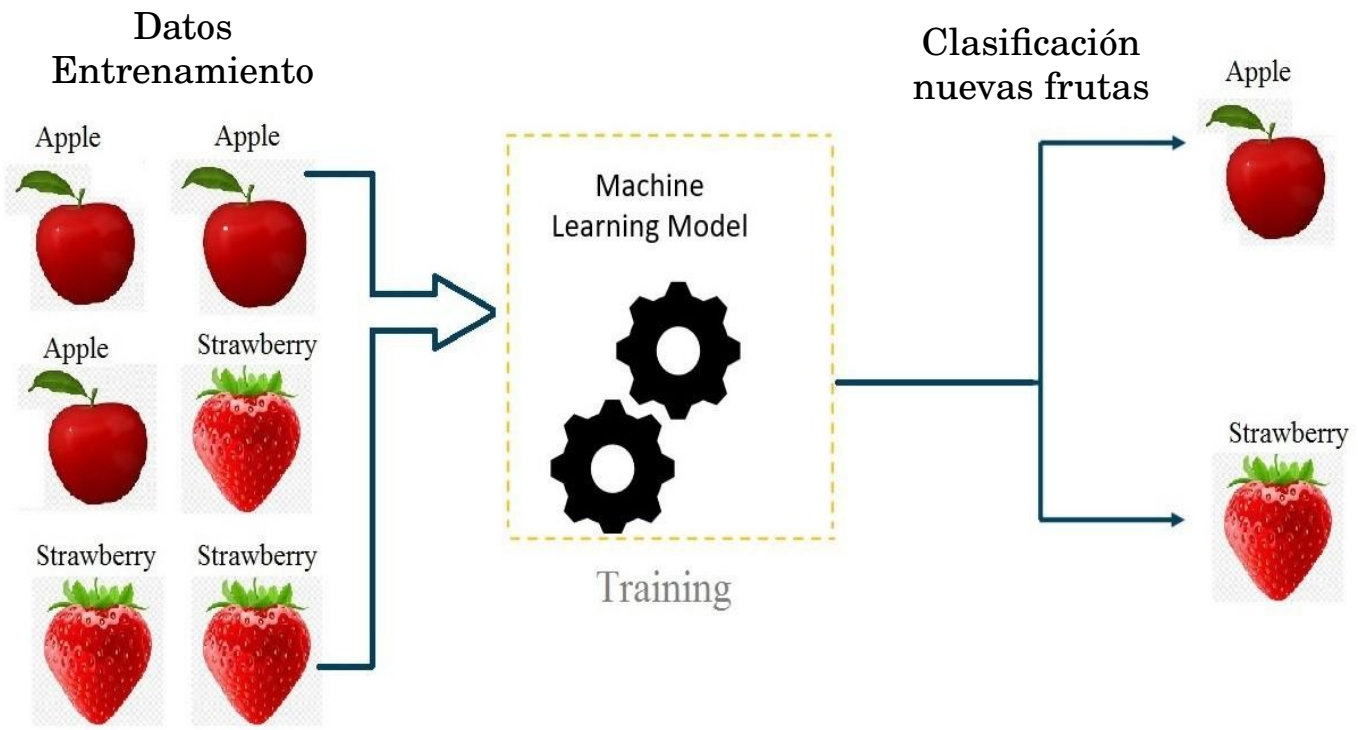
Los datos surgen por interacción con el entorno, y el aprendizaje surge gradualmente en base a una recompensa.

Aprendizaje supervisado

El aprendizaje supervisado permite que los algoritmos 'aprendan' de datos históricos/de entrenamiento y los apliquen a entradas desconocidas para obtener la salida correcta.

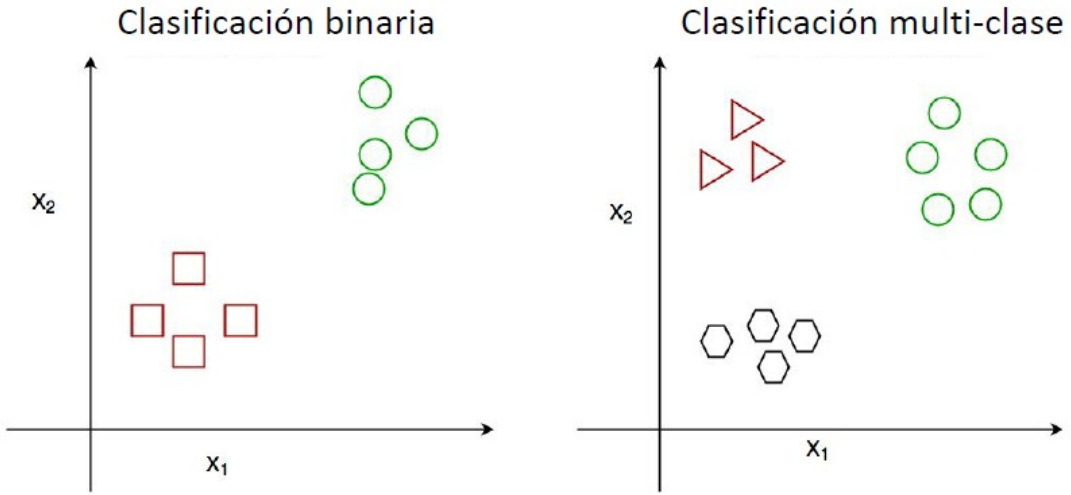
Para funcionar, el aprendizaje supervisado utiliza árboles de decisión, bosques aleatorios y Gradient Boosting Machine.

Existen dos tipos principales de aprendizaje supervisado; **clasificación y regresión**

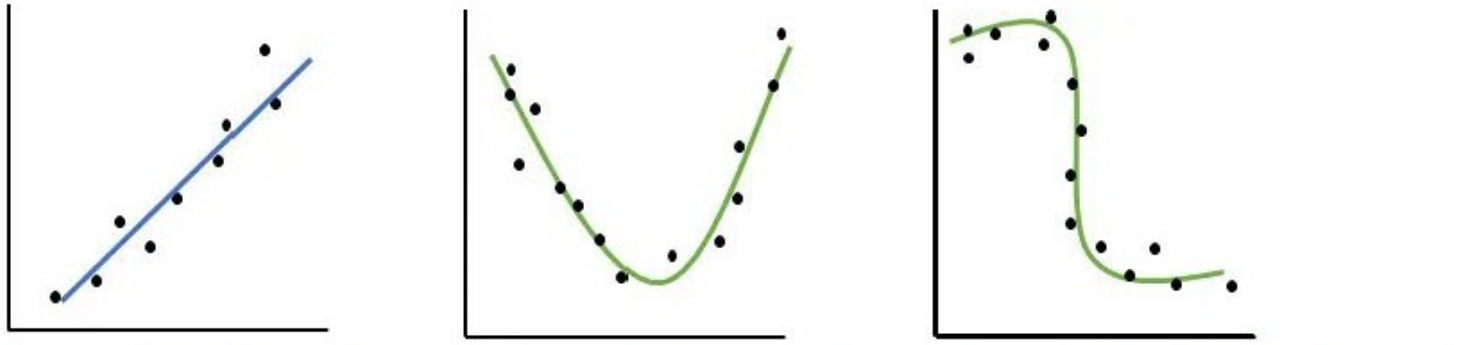


Aprendizaje supervisado

Clasificación:



Regresión:

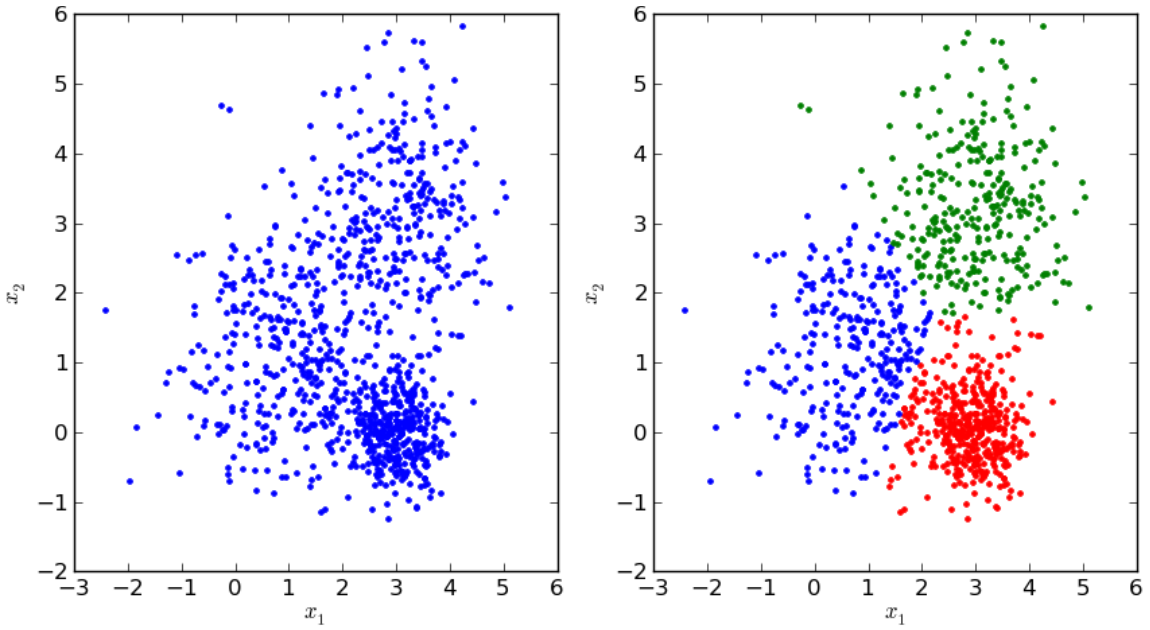


Regresión lineal

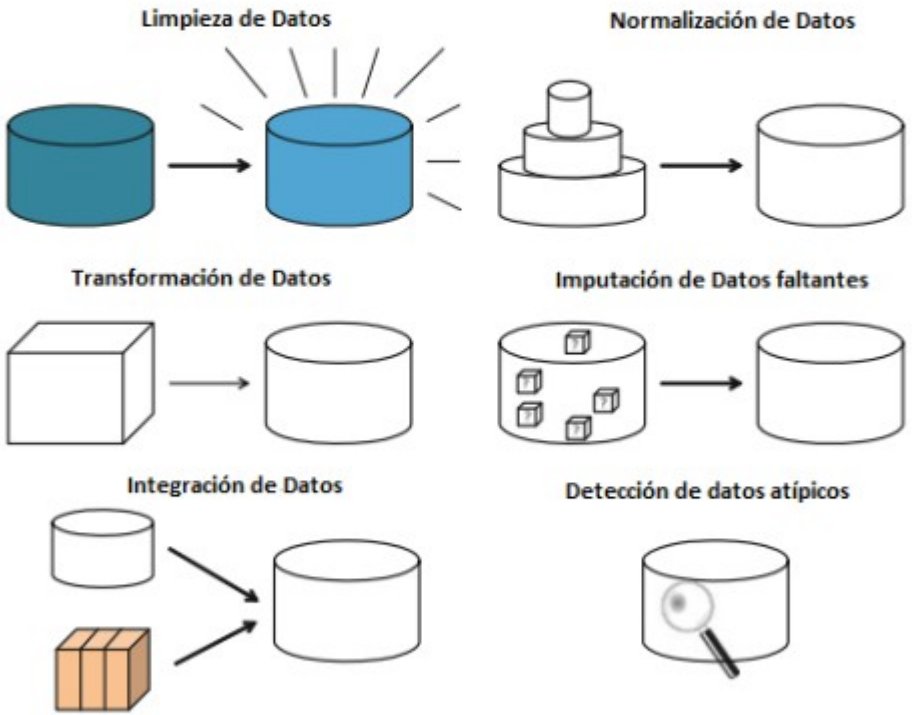
Regresión no lineal (polinomio, etc)

Aprendizaje no supervisado

El aprendizaje no supervisado tiene datos sin etiquetar que el algoritmo tiene que intentar entender por sí mismo. El objetivo es simplemente dejar que la máquina aprenda sin ayuda o indicaciones de los científicos de datos, también deberá aprender a ajustar los resultados y agrupaciones cuando haya resultados más adecuados, permitiendo que la máquina comprenda los datos y los procese como mejor le parezca. Se utiliza para explorar datos desconocidos. Puede revelar patrones que podrían haberse pasado por alto o examinar grandes conjuntos de datos que serían demasiado para que los aborde una sola persona



Preprocesamiento de Datos



El preprocesamiento de datos implica una serie de pasos necesarios para poder extraer información de ellos y construir modelos de AA que sean de utilidad.

Algunas de las tareas más comunes son:

- Normalización de Datos
- Transformación de Datos
- Imputación de Datos faltantes
- Detección de datos atípicos
- Integración de datos
- Reducción de la dimensionalidad

Limpieza de datos

- **Datos faltantes:** no siempre vamos a contar con datos en todos los atributos de nuestros datos. La ausencia de un dato puede deberse al azar o estar vinculada con alguna otra variable o hecho que no conocemos.
  - Estrategias:
    - Eliminar el dato/atributo
    - Imputar el dato faltante (intentar estimar cuál es el valor que falta a partir del resto de los datos que tengo)
    - No hacer nada (algunos modelos admiten datos faltantes)

- **Datos atípicos:** pueden deberse a error en la toma de los datos o en su procesamiento. El desafío es poder distinguir los casos en los que un dato atípico representa verdaderamente un error y los que no.
  - Estrategias:
    - A los fines prácticos, se pueden tratar de la misma manera que a los datos faltantes.

UNAHUR

Limpieza de datos

- **Transformación de datos:** muchas veces en nuestros modelos utilizamos atributos que son de distinta naturaleza (ej: variables categóricas y numéricas) o que varían en escalas muy distintas (ej: edad y altura de las personas). En estos casos, nos gustaría que nuestros atributos se "parezcan más entre sí".
  - Estandarización/Normalización:
    - Nos permite llevar a todas nuestras variables numéricas a la misma escala de variación, haciendo que sean más comparables entre sí.
    - Ayuda a evitar que atributos con mayores magnitudes tengan a su vez un mayor peso en los modelos que el resto de los atributos.
    - Métodos más usados: Min-Max, Z-Score y Decimal Scaling.

UNAHUR

Limpieza de datos

- **Transformación de datos:** muchas veces en nuestros modelos utilizamos atributos que son de distinta naturaleza (ej: variables categóricas y numéricas) o que varían en escalas muy distintas (ej: edad y altura de las personas). En estos casos, nos gustaría que nuestros atributos se "parezcan más entre sí".
  - Estandarización/Normalización:
    - Min-Max:

$$X^*_{\text{mm}} = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- Valores normalizados van de 0 a 1.
- Dominada por los valores atípicos

```
Sepal.Length

Min.    4.300
1st Qu. 5.100
Median  5.800
Mean    5.843
3rd Qu. 6.400
Max.    7.900
```

$$X_{mm} = \frac{X-4.3}{7.9-4.3}$$

Para los valores extremos es 0 y 1

Limpieza de datos

- **Transformación de datos:** muchas veces en nuestros modelos utilizamos atributos que son de distinta naturaleza (ej: variables categóricas y numéricas) o que varían en escalas muy distintas (ej: edad y altura de las personas). En estos casos, nos gustaría que nuestros atributos se "parezcan más entre sí".
  - Estandarización/Normalización:
    - Z-score:

$$Z-score = \frac{X - mean(X)}{sd(X)}$$

Útil cuando…

- el verdadero mínimo y máximo son desconocidos
- hay valores atípicos que dominan la normalización min-max. Acá se puede usar la mediana

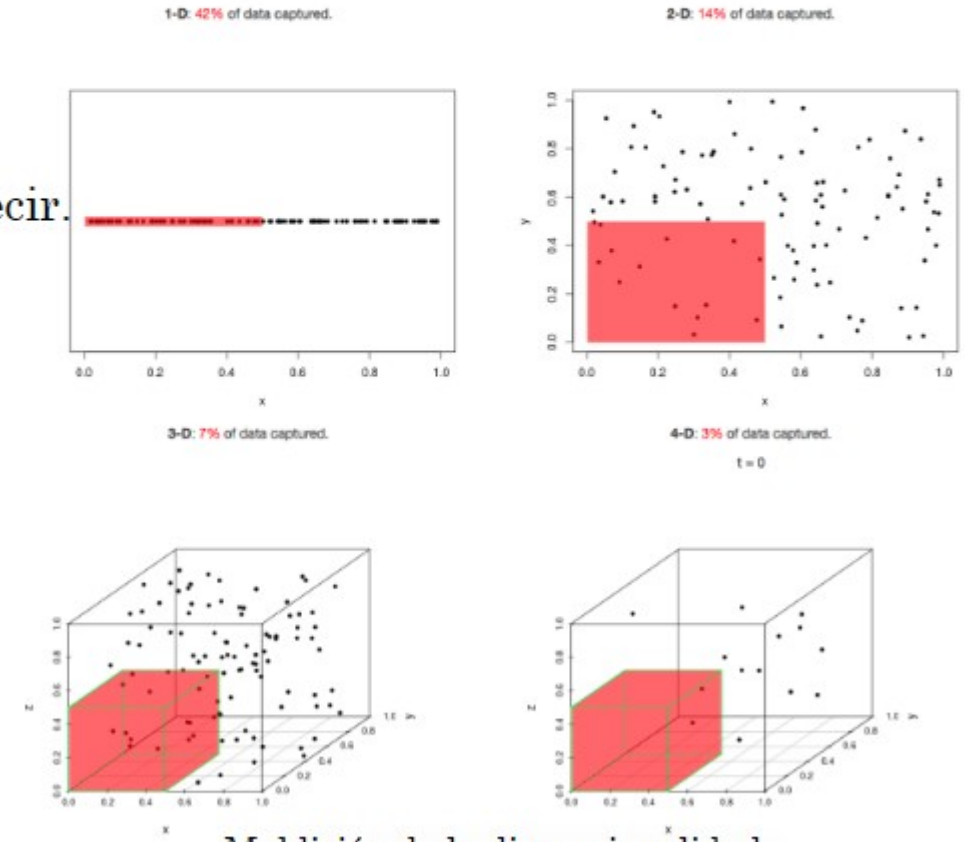
```
Sepal.Length

Min.    4.300
1st Qu. 5.100
Median  5.800
Mean    5.843
3rd Qu. 6.400
Max.    7.900
```

$$Z-score = \frac{4.3 - 5.843}{0.828} = -1,863$$

$$Z-score = \frac{7.9 - 5.843}{0.828} = 2,484$$

Limpieza de datos

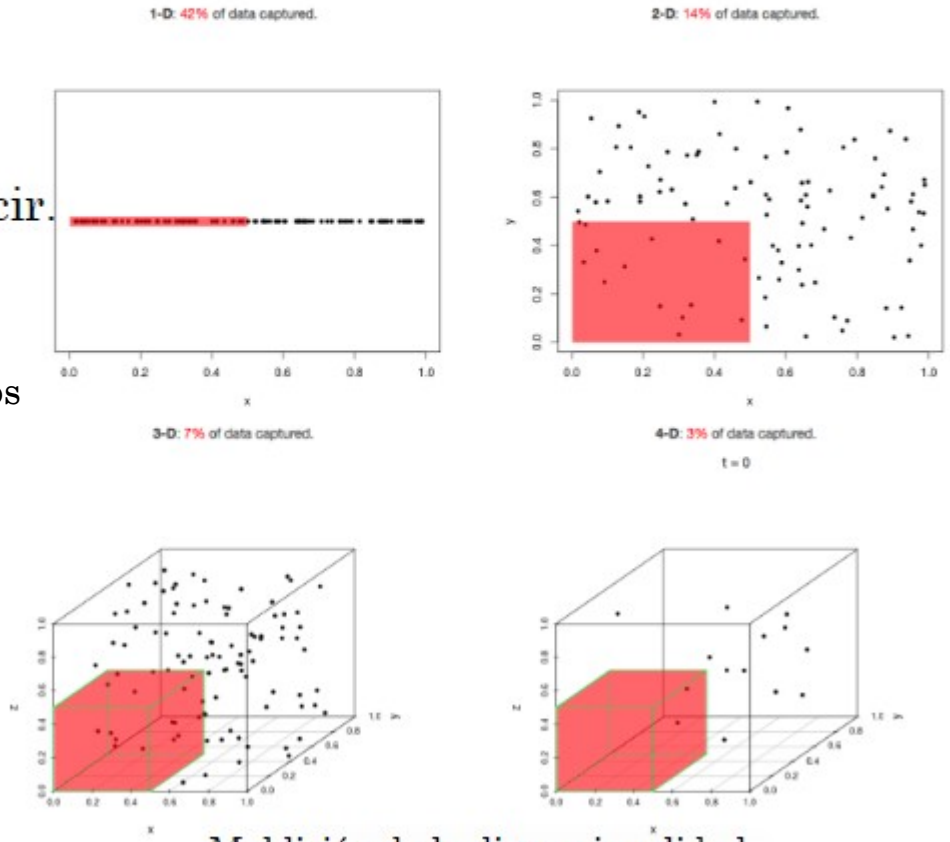
- **<u>Reducción de la dimensionalidad;</u>** muchas veces en nuestras bases de datos tenemos mucha redundancia que afecta negativamente la performance de los modelos que queramos hacer.
  - Presencia de registros/filas repetidas.
  - Atributos altamente correlacionados.
  - Atributos con muy poca varianza.
  - Atributos poco relacionados con la variable a predecir.



Maldición de la dimensionalidad

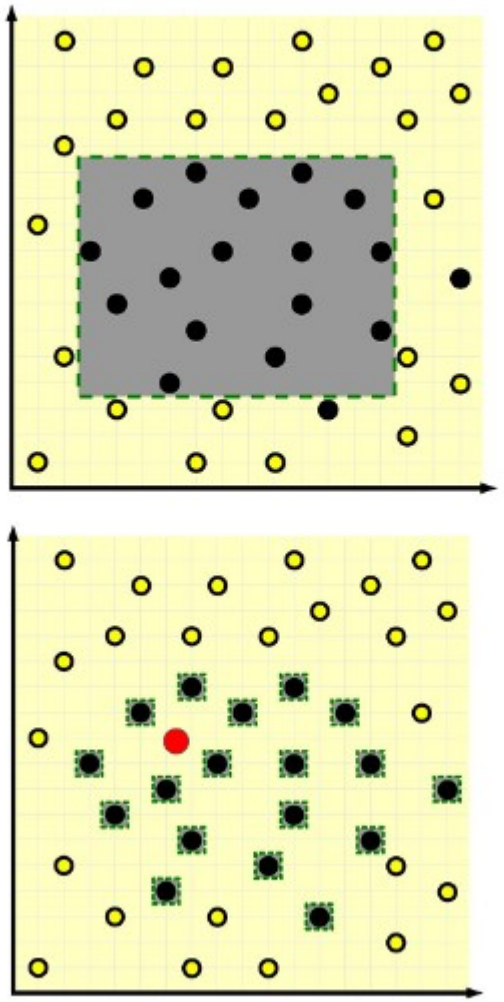
Limpieza de datos

- **Reducción de la dimensionalidad;** muchas veces en nuestras bases de datos tenemos mucha redundancia que afecta negativamente la performance de los modelos que queramos hacer.
  - Presencia de registros/filas repetidas.
  - Atributos altamente correlacionados.
  - Atributos con muy poca varianza.
  - Atributos poco relacionados con la variable a predecir.
  - → **Modelos no supervisados!** (PCA)
    - Ayudan a reducir la dimensionalidad conservando (casi) toda la información de nuestros datos.
    - Muy útiles para la representación gráfica de nuestros datos de alta dimensionalidad



Maldición de la dimensionalidad

Complejidad de los Modelos



Diferencia entre Aprender y Memorizar

Puedo hacer un modelo que se dedique a memorizar los datos que le doy para entrenar. Si hago eso podría alcanzar un accuracy de 1 y… tendría el modelo perfecto?? NO!

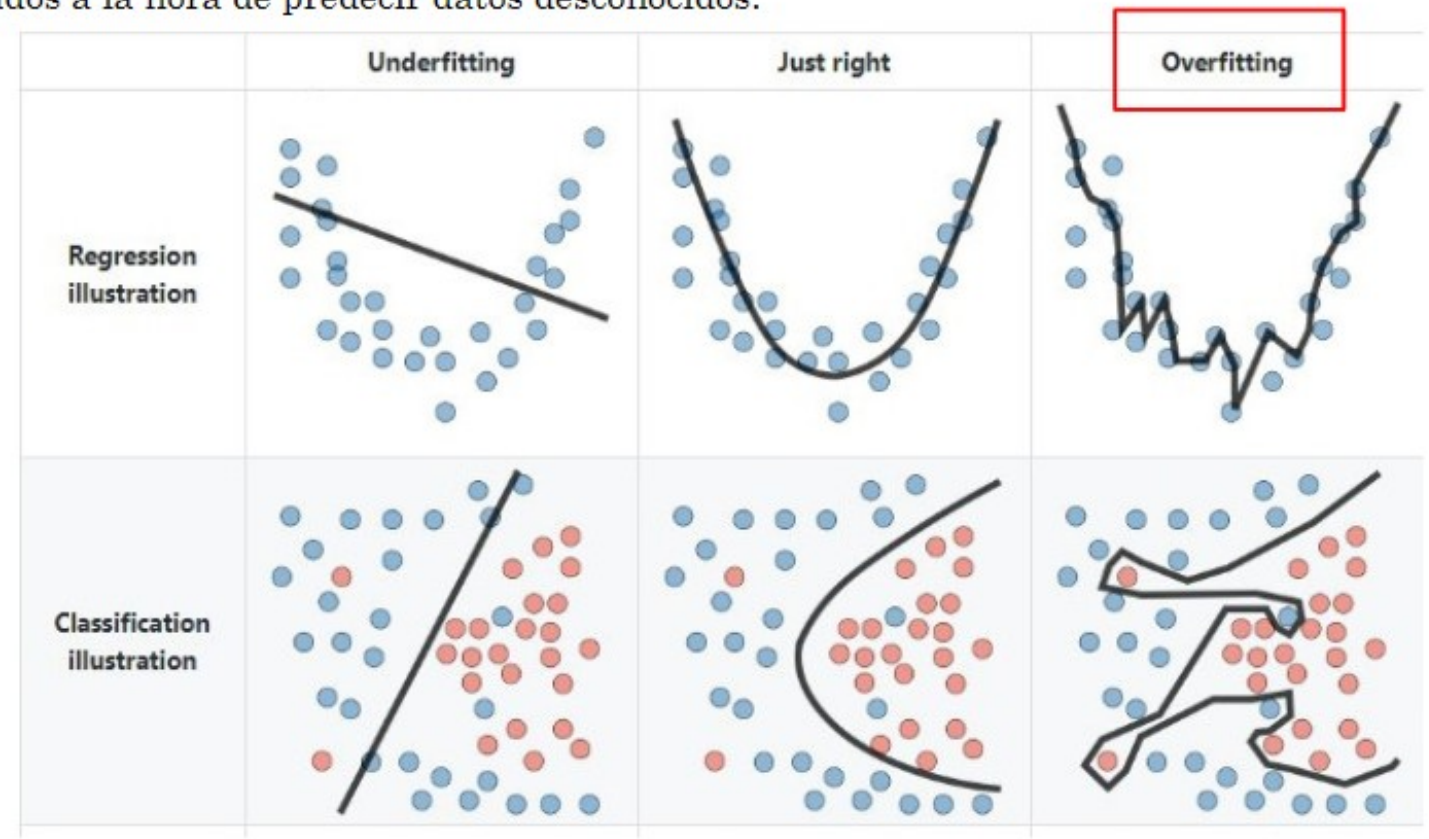
Para aprender las personas (y las máquinas) necesitan poder encontrar patrones que les permitan generalizar sus afirmaciones. En caso contrario, mi modelo no me va a servir cuando quiera predecir un punto que nunca vió.

En igualdad de condiciones, elegir la explicación más simple.

Navaja de Ockham

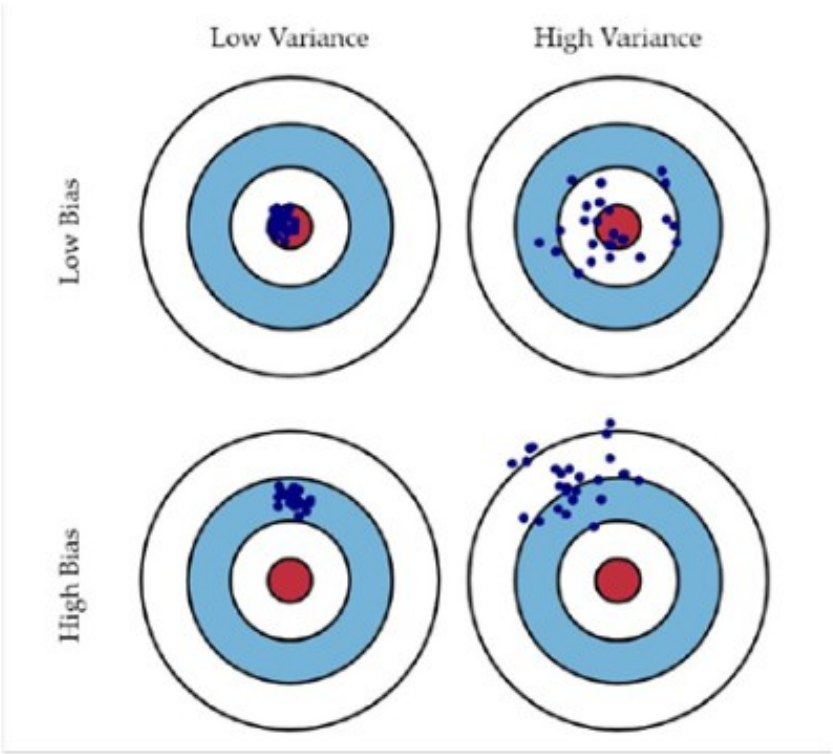
Problema: relación sesgo-varianza

Nuestra misión es construir un modelo que no solo logre representar bien los datos que tengo, sino también, que permita predecir nuevos datos que nunca vió (diferencia entre memorizar y aprender). Cuanto más complejo sea nuestro modelo (más parámetros tenga) posiblemente le vaya mejor explicando los datos que tengo pero le va a tener peores resultados a la hora de predecir datos desconocidos.



Problema: relación sesgo-varianza

Nuestra misión es construir un modelo que no solo logre representar bien los datos que tengo, sino también, que permita predecir nuevos datos que nunca vió (diferencia entre memorizar y aprender). Cuanto más complejo sea nuestro modelo (más parámetros tenga) posiblemente le vaya mejor explicando los datos que tengo pero le va a tener peores resultados a la hora de predecir datos desconocidos.



El **sesgo** o Bias en un modelo de machine learning es un tipo de error que indica la diferencia que existe entre la predicción del modelo y el valor actual. Si el modelo tiene un Bias alto significa que le presta poca atención a los datos y sobre simplifica el modelo.

La **varianza** también también es parte del error en un modelo de machine learning. Este error lo podemos entender como que tan sensible es nuestro modelo a los datos. Si tenemos exceso de sensibilidad el modelo puede creer ver patrones que realmente no existen ahí. Si tenemos un valor de varianza alto esto significa que el modelo le presta mucha atención a los datos de entrenamiento y no va a generalizar bien en datos que no ha visto.

$$Error(x) = Bias^2 + Variance + ErrorIrreductible$$